The American Economic Review

ARTICLES

[illegible]

[illegible] — Optimal Taxation and Public Production I

[illegible]

[illegible]

[illegible] — On the Theory of the Competitive Firm Under Price Uncertainty

[illegible] — The Effect of [illegible] Consumption and Trade: A [illegible] Analysis

[illegible] — A General Disequilibrium Model of Income and Employment

[illegible] — A Test for Stability [illegible] and Application to [illegible]

[illegible]

[illegible] — Determinants of the Commodity Structure of U.S. Trade

[illegible]

VOL. [illegible] MARCH 1971

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, and *Directory* and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September and December.

• Membership dues of the Association are $20.00 a year, which includes a year's subscription to both the *American Economic Review* and the *Journal of Economic Literature*. Subscriptions by nonmembers are $20.00 a year, and only subscriptions to both publications will be accepted. Single copies of the *Review* and *Journal* are $4.00 each. Each order for copies of either publication must also include a $.50 per order service charge. Orders should be sent to the Secretary's office, Nashville, Tennessee.

• Correspondence relating to the *Paper and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address may be sent to the secretary, Rendigs Fels, 1313 21st Avenue, South, Nashville, Tennessee 37212. To be effective, notice of change of address must reach the secretary by the 1st of the month previous to the month of publication.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

THE AMERICAN ECONOMIC REVIEW

March 1971

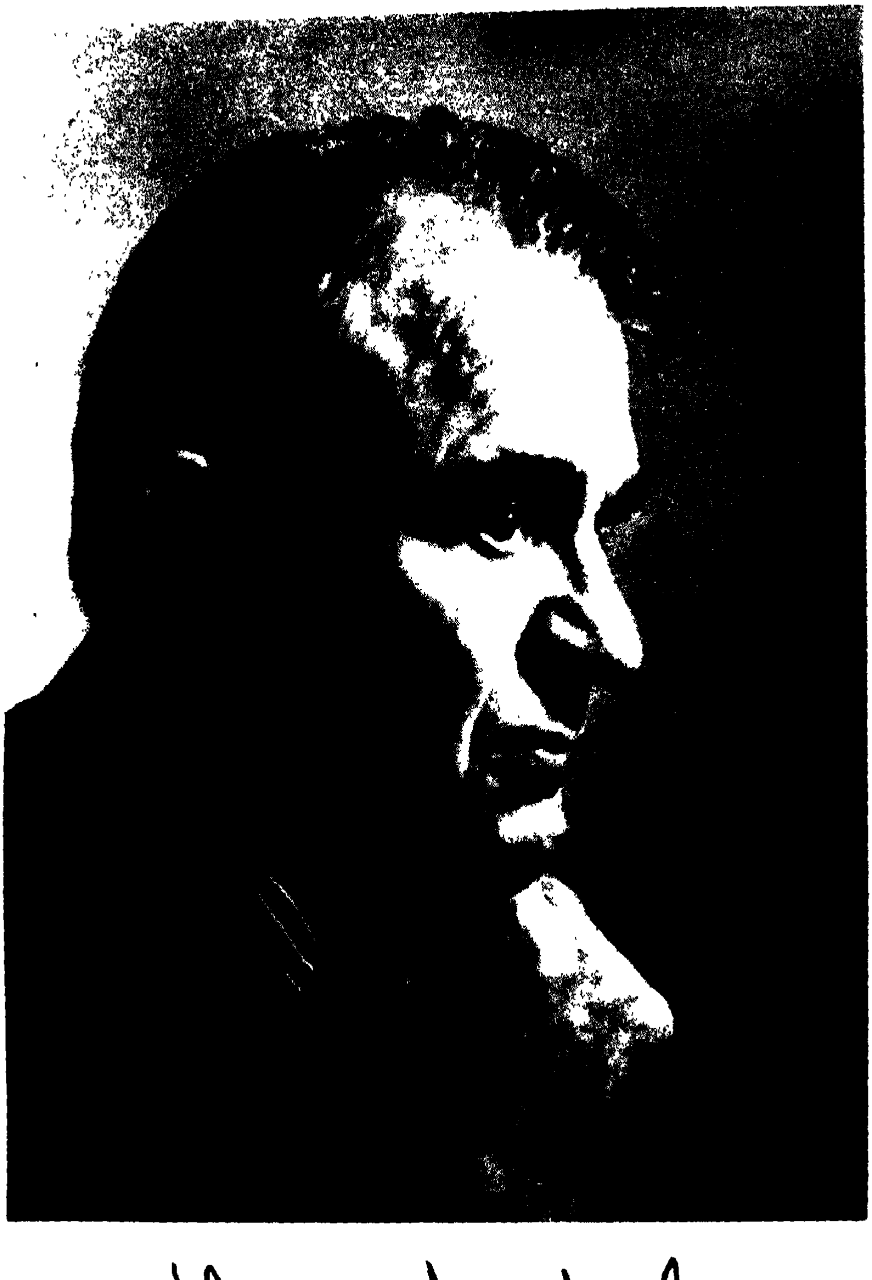
VOLUME LXI, NUMBER 1

Articles

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

Communications

Number 72 of a series of photographs of past presidents of the Association



Wassily Leontief

Theoretical Assumptions and Nonobserved Facts

By Wassily Leontief*

Economics today rides the crest of intellectual respectability and popular acclaim. The serious attention with which our pronouncements are received by the general public, hard-bitten politicians, and even skeptical businessmen is second only to that which was given to physicists and space experts a few years ago when the round trip to the moon seemed to be our only truly national goal. The flow of learned articles, monographs, and textbooks is swelling like a tidal wave; *Econometrica*, the leading journal in the field of mathematical economics, has just stepped up its publication schedule from four to six issues per annum.

And yet an uneasy feeling about the present state of our discipline has been growing in some of us who have watched its unprecedented development over the last three decades. This concern seems to be shared even by those who are themselves contributing successfully to the present boom. They play the game with professional skill but have serious doubts about its rules.

Much of current academic teaching and research has been criticized for its lack of relevance, that is, of immediate practical impact. In a nearly instant response to this criticism, research projects, seminars and undergraduate courses have been set up on poverty, on city and small town slums, on pure water and fresh air. In an almost Pavlovian reflex, whenever a new complaint is raised, President Nixon appoints a commission and the university announces a new course. Far be it from me to argue that the fire should not be shifted when the target moves. The trouble is caused, however, not by an inadequate selection of targets, but rather by our inability to hit squarely any one of them. The uneasiness of which I spoke before is caused not by the *irrelevance* of the practical problems to which present day economists address their efforts, but rather by the palpable *inadequacy* of the scientific means with which they try to solve them.

If this simply were a sign of the overly high aspiration level of a fast developing discipline, such a discrepancy between ends and means should cause no worry. But I submit that the consistently indifferent performance in practical applications is in fact a symptom of a fundamental imbalance in the present state of our discipline. The weak and all too slowly growing empirical foundation clearly cannot support the proliferating superstructure of pure, or should I say, speculative economic theory.

Much is being made of the widespread, nearly mandatory use by modern economic theorists of mathematics. To the extent to which the economic phenomena possess observable quantitative dimensions, this is indisputably a major forward step. Unfortunately, any one capable of learning elementary, or preferably advanced calculus and algebra, and acquiring acquaintance with the specialized terminology of economics can set himself up as a theorist. Uncritical enthusiasm for mathematical formulation tends often to con-

* Presidential address delivered at the eighty-third meeting of The American Economic Association, Detroit, Michigan, December 29, 1970.

ceal the ephemeral substantive content of the argument behind the formidable front of algebraic signs.

Professional journals have opened wide their pages to papers written in mathematical language; colleges train aspiring young economists to use this language; graduate schools require its knowledge and reward its use. The mathematical model-building industry has grown into one of the most prestigious, possibly the most prestigious branch of economics. Construction of a typical theoretical model can be handled now as a routine assembly job. All principal components such as production functions, consumption and utility functions come in several standard types; so does the optional equipment as, for example, "factor augmentation"—to take care of technological change. This particular device is, incidentally, available in a simple exponential design or with a special automatic regulator known as the "Kennedy function." Any model can be modernized with the help of special attachments. One popular way to upgrade a simple one-sector model is to bring it out in a two-sector version or even in a still more impressive form of the "*n*-sector," that is, many-sector class.

In the presentation of a new model, attention nowadays is usually centered on a step-by-step derivation of its formal properties. But if the author—or at least the referee who recommended the manuscript for publication—is technically competent, such mathematical manipulations, however long and intricate, can even without further checking be accepted as correct. Nevertheless, they are usually spelled out at great length. By the time it comes to interpretation of the substantive *conclusions*, the assumptions on which the model has been based are easily forgotten. But it is precisely the empirical validity of these *assumptions* on which the usefulness of the entire exercise depends.

What is really needed, in most cases, is a very difficult and seldom very neat assessment and verification of these assumptions in terms of observed facts. Here mathematics cannot help and because of this, the interest and enthusiasm of the model builder suddenly begins to flag: "If you do not like my set of assumptions, give me another and I will gladly make you another model; have your pick."

Policy oriented models, in contrast to purely descriptive ones, are gaining favor, however nonoperational they may be. This, I submit, is in part because the choice of the final policy objectives—the selection and justification of the shape of the so-called objective function—is, and rightly so, considered based on normative judgment, not on factual analysis. Thus, the model builder can secure at least some convenient assumptions without running the risk of being asked to justify them on empirical grounds.

To sum up with the words of a recent president of the Econometric Society, " . . . the achievements of economic theory in the last two decades are both impressive and in many ways beautiful. But it cannot be denied that there is something scandalous in the spectacle of so many people refining the analysis of economic states which they give no reason to suppose will ever, or have ever, come about. . . . It is an unsatisfactory and slightly dishonest state of affairs."

But shouldn't this harsh judgment be suspended in the face of the impressive volume of econometric work? The answer is decidedly no. This work can be in general characterized as an attempt to compensate for the glaring weakness of the data base available to us by the widest possible use of more and more sophisticated statistical techniques. Alongside the mounting pile of elaborate theoretical models we see a fast-growing stock of equally intricate statistical tools. These

are intended to stretch to the limit the meager supply of facts.

Since, as I said before, the publishers' referees do a competent job, most model-testing kits described in professional journals are internally consistent. However, like the economic models they are supposed to implement, the validity of these statistical tools depends itself on the acceptance of certain convenient assumptions pertaining to stochastic properties of the phenomena which the particular models are intended to explain; assumptions that can be seldom verified.

In no other field of empirical inquiry has so massive and sophisticated a statistical machinery been used with such indifferent results. Nevertheless, theorists continue to turn out model after model and mathematical statisticians to devise complicated procedures one after another. Most of these are relegated to the stockpile without any practical application or after only a perfunctory demonstration exercise. Even those used for a while soon fall out of favor, not because the methods that supersede them perform better, but because they are new and different.

Continued preoccupation with imaginary, hypothetical, rather than with observable reality has gradually led to a distortion of the informal valuation scale used in our academic community to assess and to rank the scientific performance of its members. Empirical analysis, according to this scale, gets a lower rating than formal mathematical reasoning. Devising a new statistical procedure, however tenuous, that makes it possible to squeeze out one more unknown parameter from a given set of data, is judged a greater scientific achievement than the successful search for additional information that would permit us to measure the magnitude of the same parameter in a less ingenious, but more reliable way. This despite the fact that in all too many instances sophisticated statistical analysis is performed on a set of data whose exact meaning and validity are unknown to the author or rather so well known to him that at the very end he warns the reader not to take the material conclusions of the entire "exercise" seriously.

A natural Darwinian feedback operating through selection of academic personnel contributes greatly to the perpetuation of this state of affairs. The scoring system that governs the distribution of rewards must naturally affect the make-up of the competing teams. Thus, it is not surprising that the younger economists, particularly those engaged in teaching and in academic research, seem by now quite content with a situation in which they can demonstrate their prowess (and incidentally, advance their careers) by building more and more complicated mathematical models and devising more and more sophisticated methods of statistical inference without ever engaging in empirical research. Complaints about the lack of indispensable primary data are heard from time to time, but they don't sound very urgent. The feeling of dissatisfaction with the present state of our discipline which prompts me to speak out so bluntly seems, alas, to be shared by relatively few. Yet even those few who do share it feel they can do little to improve the situation. How could they?

In contrast to most physical sciences, we study a system that is not only exceedingly complex but is also in a state of constant flux. I have in mind not the obvious change in the variables, such as outputs, prices or levels of employment, that our equations are supposed to explain, but the basic structural relationships described by the form and the parameters of these equations. In order to know what the shape of these structural relationships actually are at any given time, we have to keep them under continuous surveillance.

By sinking the foundations of our ana-

lytical system deeper and deeper, by reducing, for example, cost functions to production functions and the production functions to some still more basic relationships eventually capable of explaining the technological change itself, we should be able to reduce this drift. It would, nevertheless, be quite unrealistic to expect to reach, in this way, the bedrock of invariant structural relationships (measurable parameters) which, once having been observed and described, could be used year after year, decade after decade, without revisions based on repeated observation.

On the relatively shallow level where the empirically implemented economic analysis now operates even the more invariant of the structural relationships, in terms of which the system is described, change rapidly. Without a constant inflow of new data the existing stock of factual information becomes obsolete very soon. What a contrast with physics, biology or even psychology where the magnitude of most parameters is practically constant and where critical experiments and measurements don't have to be repeated every year!

Just to keep up our very modest current capabilities we have to maintain a steady flow of new data. A progressive expansion of these capabilities would be out of the question without a continuous and rapid rise of this flow. Moreover, the new, additional data in many instances will have to be qualitatively different from those provided hitherto.

To deepen the foundation of our analytical system it will be necessary to reach unhesitatingly beyond the limits of the domain of economic phenomena as it has been staked out up to now. The pursuit of a more fundamental understanding of the process of production inevitably leads into the area of engineering sciences. To penetrate below the skin-thin surface of conventional consumption functions, it will be necessary to develop a systematic study of the structural characteristics and of the functioning of households, an area in which description and analysis of social, anthropological and demographic factors must obviously occupy the center of the stage.

Establishment of systematic cooperative relationships across the traditional frontiers now separating economics from these adjoining fields is hampered by the sense of self-sufficiency resulting from what I have already characterized as undue reliance on indirect statistical inference as the principal method of empirical research. As theorists, we construct systems in which prices, outputs, rates of saving and investment, etc., are explained in terms of production functions, consumption functions and other structural relationships whose parameters are assumed, at least for arguments' sake, to be known. As econometricians, engaged in what passes for empirical research, we do not try, however, to ascertain the actual shapes of these functions and to measure the magnitudes of these parameters by turning up new factual information. We make an about face and rely on indirect statistical inference to derive the unknown structural relationships from the observed magnitudes of prices, outputs and other variables that, in our role as theoreticians, we treated as unknowns.

Formally, nothing is, of course, wrong with such an apparently circular procedure. Moreover, the model builder in erecting his hypothetical structures is free to take into account all possible kinds of factual knowledge and the econometrician in principle, at least, can introduce in the estimating procedure any amount of what is usually referred to as "exogenous" information before he feeds his pro-

grammed tape into the computer. Such options are exercised rarely and when they are, usually in a casual way.

The same well-known sets of figures are used again and again in all possible combinations to pit different theoretical models against each other in formal statistical combat. For obvious reasons a decision is reached in most cases not by a knock-out, but by a few points. The orderly and systematic nature of the entire procedure generates a feeling of comfortable self-sufficiency.

This complacent feeling, as I said before, discourages venturesome attempts to widen and to deepen the empirical foundations of economic analysis, particularly those attempts that would involve crossing the conventional lines separating ours from the adjoining fields.

True advance can be achieved only through an iterative process in which improved theoretical formulation raises new empirical questions and the answers to these questions, in their turn, lead to new theoretical insights. The "givens" of today become the "unknowns" that will have to be explained tomorrow. This, incidentally, makes untenable the admittedly convenient methodological position according to which a theorist does not need to verify directly the factual assumptions on which he chooses to base his deductive arguments, provided his empirical conclusions seem to be correct. The prevalence of such a point of view is, to a large extent, responsible for the state of splendid isolation in which our discipline nowadays finds itself.

An exceptional example of a healthy balance between theoretical and empirical analysis and of the readiness of professional economists to cooperate with experts in the neighboring disciplines is offered by Agricultural Economics as it developed in this country over the last fifty years. A unique combination of social and political forces has secured for this area unusually strong organizational and generous financial support. Official agricultural statistics are more complete, reliable, and systematic than those pertaining to any other major sector of our economy. Close collaboration with agronomists provides agricultural economists with direct access to information of a technological kind. When they speak of crop rotation, fertilizers, or alternative harvesting techniques, they usually know, sometimes from personal experience, what they are talking about. Preoccupation with the standard of living of the rural population has led agricultural economists into collaboration with home economists and sociologists, that is, with social scientists of the "softer" kind. While centering their interest on only one part of the economic system, agricultural economists demonstrated the effectiveness of a systematic combination of theoretical approach with detailed factual analysis. They also were the first among economists to make use of the advanced methods of mathematical statistics. However, in their hands, statistical inference became a complement to, not a substitute for, empirical research.

The shift from casual empiricism that dominates much of today's econometric work to systematic large-scale factual analysis will not be easy. To start with, it will require a sharp increase in the annual appropriation for Federal Statistical Agencies. The quality of government statistics has, of course, been steadily improving. The coverage, however, does not keep up with the growing complexity of our social and economic system and our capability of handling larger and larger data flows.

The spectacular advances in computer technology increased the economists' potential ability to make effective analytical use of large sets of detailed data. The time

is past when the best that could be done with large sets of variables was to reduce their number by averaging them out or what is essentially the same, combining them into broad aggregates; now we can manipulate complicated analytical systems without suppressing the identity of their individual elements. There is a certain irony in the fact that, next to the fast-growing service industries, the areas whose coverage by the Census is particularly deficient are the operations of government agencies, both federal and local.

To place all or even the major responsibility for the collection of economic data in the hands of one central organization would be a mistake. The prevailing decentralized approach that permits and encourages a great number of government agencies, non-profit institutions and private businesses engaged in data gathering activities acquitted itself very well. Better information means more detailed information and detailed specialized information can be best collected by those immediately concerned with a particular field. What is, however, urgently needed is the establishment, maintenance and enforcement of coordinated uniform classification systems by all agencies, private as well as public, involved in this work. Incompatible data are useless data. How far from a tolerable, not to say, ideal state our present economic statistics are in this respect, can be judged by the fact that because of differences in classification, domestic output data cannot be compared, for many goods, with the corresponding export and import figures. Neither can the official employment statistics be related without laborious adjustments to output data, industry by industry. An unreasonably high proportion of material and intellectual resources devoted to statistical work is now spent not on the collection of primary information but on a frustrating and wasteful struggle with incongruous definitions and irreconcilable classifications.

Without invoking a misplaced methodological analogy, the task of securing a massive flow of primary economic data can be compared to that of providing the high energy physicists with a gigantic accelerator. The scientists have their machines while the economists are still waiting for their data. In our case not only must the society be willing to provide year after year the millions of dollars required for maintenance of a vast statistical machine, but a large number of citizens must be prepared to play, at least, a passive and occasionally even an active part in actual fact-finding operations. It is as if the electrons and protons had to be persuaded to cooperate with the physicist.

The average American does not seem to object to being interviewed, polled, and surveyed. Curiosity, the desire to find out how the economic system (in which most of us are small gears, and some, big wheels) works might in many instances provide sufficient inducement for cooperation of this kind.

One runs up, of course, occasionally against the attitude that "what you don't know can't hurt you" and that knowledge might be dangerous: it may generate a desire to tinker with the system. The experience of these years seems, however, to have convinced not only most economists—with a few notable exceptions—but also the public at large that a lack of economic knowledge can hurt badly. Our free enterprise system has rightly been compared to a gigantic computing machine capable of solving its own problems automatically. But any one who has had some practical experience with large computers knows that they do break down and can't operate unattended. To keep the automatic, or rather the semi-automatic, engine of our economy in good working order we must not only understand the general

principles on which it operates, but also be acquainted with the details of its actual design.

A new element has entered the picture in recent years—the adoption of methods of modern economic analysis by private business. Corporate support of economic research goes as far back as the early 1920's when Wesley Mitchell founded the National Bureau. However, it is not this concern for broad issues of public policies or even the general interest in economic growth and business fluctuations that I have in mind, but rather, the fast-spreading use of advanced methods of Operations Research and of so-called Systems' Analysis. Some of the standard concepts and analytical devices of economic theory first found their way into the curricula of our business schools and soon after that, sophisticated management began to put them into practice. While academic theorists are content with the formulation of general principles, corporate operations researchers and practical systems' analysts have to answer questions pertaining to specific real situations. Demand for economic data to be used in practical business planning is growing at an accelerated pace. It is a high quality demand: business users in most instances possess first-hand technical knowledge of the area to which the data they ask for refer. Moreover, this demand is usually "effective." Profit-making business is willing and able to pay the costs of gathering the information it wants to have. This raises the thorny question of public access to privately collected data and of the proper division of labor and cooperation between government and business in that fast-expanding field. Under the inexorable pressure of rising practical demand, these problems will be solved in one way or another. Our economy will be surveyed and mapped in all its many dimensions on a larger and larger scale.

Economists should be prepared to take a leading role in shaping this major social enterprise not as someone else's spokesmen and advisers, but on their own behalf. They have failed to do this up to now. The Conference of Federal Statistics Users organized several years ago had business, labor, and many other groups represented among its members, but not economists as such. How can we expect our needs to be satisfied if our voices are not heard?

We, I mean the academic economists, are ready to expound, to any one ready to lend an ear, our views on problems of public policy: give advice on the best ways to maintain full employment, to fight inflation, to foster economic growth. We should be equally prepared to share with the wider public the hopes and disappointments which accompany the advance of our own often desperately difficult, but always exciting intellectual enterprise. This public has amply demonstrated its readiness to back the pursuit of knowledge. It will lend its generous support to our venture too, if we take the trouble to explain what it is all about.

REFERENCE

F. H. Hahn, "Some Adjustment Problems," *Econometrica*, Jan. 1970, *38*, 1–2.

Optimal Taxation and Public Production I: Production Efficiency

By PETER A. DIAMOND AND JAMES A. MIRRLEES*

Theories of optimal production in a planned economy have usually assumed that the tax system can allow the government to achieve any desired redistribution of property.[1] On the other hand, some recent discussions of public investment criteria have tended to ignore taxation as a complementary method of controlling the economy.[2] Although lump sum transfers of the kind required for full optimality[3] are not feasible today, commodity and income taxes can certainly be used to increase welfare.[4] We shall therefore examine the maximization of social welfare using both taxes and public production as control variables. In doing so, we intend to bring together the theories of taxation, public investment, and welfare economics.

There are two main results of the study: the demonstration of the desirability of aggregate production efficiency in a wide variety of circumstances provided that taxes are set at the optimal level; and an examination of that optimal tax structure. It is widely known that aggregate production efficiency is desired as one part of achieving a Pareto optimum. It is also widely known that when the desired Pareto optimum cannot be achieved, aggregate production efficiency may not be desirable. Our conclusion differs from these results in that production efficiency is desirable although a full Pareto optimum is not achieved. In the optimum position, the presence of commodity taxes implies that marginal rates of substitution are not equal to marginal rates of transformation. Furthermore, the absence of lump sum taxes implies that the income distribution is not the best that can be conceived. Yet, the presence of optimal commodity taxes will be shown to imply the desirability of aggregate production efficiency.

This result is similar to that derived by Marcel Boiteux, although he considered an economy where lump sum redistributions of income were possible. Boiteux also examined the optimal tax structure that was necessary for this result. The optimal tax structure for the case of a single consumer (or equivalently with lump sum redistribution) has also been examined by Frank

* The authors are at Massachusetts Institute of Technology and Nuffield College, Oxford, respectively. During some of the work, Diamond was at Churchill College, Cambridge and Nuffield College, Oxford and Mirrlees was at M.I.T. Earlier versions of this paper were given at Econometric Society winter meetings at Washington and Blaricum, 1967, at the University Social Science Council Conference, Kampala, Uganda, December 1968, and to the Game Theory and Mathematical Economics Seminar, Hebrew University, Jerusalem. The authors wish to thank M.A.H. Dempster, D. K. Foley, P. A. Samuelson, K. Shell, and participants in these seminars for helpful discussions on this subject, and referees for valuable comments. Diamond was supported in part by the National Science Foundation under grant GS 1585. The authors bear sole responsibility for opinions and errors.

[1] For a discussion of this literature, see Abram Bergson.

[2] For a survey of this literature, see Alan Prest and Ralph Turvey.

[3] We wish to distinguish here between lump sum taxes, which may vary from individual to individual while being unaffected by the individual's behavior, and poll taxes which are the same for all individuals, or perhaps for all individuals within several large groups, distinguished perhaps by age, sex, or region.

[4] For another study of the general equilibrium impact of taxation, which does not explore the optimality question, see Gerard Debreu (1954).

Ramsey and Paul Samuelson.[5] Our results move beyond theirs in considering the problem of income redistribution together with that of raising revenue. Even in the absence of government revenue requirements, if lump sum redistribution is impossible, the government will want to use its excise tax powers to improve income distribution. It will subsidize and tax different goods so as to alter individual real incomes. The optimal redistribution by this method occurs when there is a balance between the equity improvements and the efficiency losses from further taxation.

The general situation we want to discuss is an economy in which there are many consumers, public and private production, public consumption, and many different kinds of feasible tax instruments. We think that it is easier to understand the problem if we present the analysis first for a single consumer, no public consumption, and only commodity taxation, although this case has little intrinsic interest. The main point of the paper is that the analysis of this special case carries over in the main to the general case.

The first two sections are devoted to this special case. In the first, the situation is portrayed geometrically (for a two-commodity world with no private production); in the second, production efficiency and conditions for the optimal taxes are derived by application of the calculus. The use of the calculus here and elsewhere is not perfectly rigorous for the usual reasons. These issues are taken up in Section IV. In the third section, we extend the analysis of production to an economy with many consumers, elucidating precise conditions under which production efficiency is desirable (and presenting certain exceptions).

[5] For a detailed history of analysis of this problem, see William Baumol and David Bradford. A summary and discussion of the work of Boiteux has been given by Jacques Drèze.

Section IV provides a rigorous statement of the theorems. In the fifth section, we discuss briefly certain applications and extensions of the basic efficiency result.

A following paper, referred to here as Diamond-Mirrlees II, will appear in the June 1971 *Review*. In it we will examine the optimality rules for commodity taxes, for other taxes including income taxes, and for public consumption. We will also give a rigorous statement of conditions under which the first-order conditions obtained (heuristically) below are indeed necessary conditions.

I. One-Consumer Economy—Geometric Analysis

We begin by considering an economy with a single, price-taking consumer and two commodities. We assume, for the moment, that all production possibilities are controlled by the government. While there is no scope for redistribution of income in this economy, the government might need to raise revenue to cover losses if there are increasing returns to scale or if there are fixed expenditures (such as defense) and constant returns to scale. Alternatively, the technology might exhibit decreasing returns to scale, facing the government with the problem of disposing of a surplus if all transactions are carried out at market prices. The optimal solution to either raising or disposing of revenue is well known. A poll tax or subsidy, as the case may be, will permit the hiring of the needed resources and permit the economy to achieve a Pareto optimum, which, in a one-consumer economy, is equivalent to the maximization of the consumer's utility. While this is a reasonable possibility in a one-consumer economy, lump sum taxes varying from individual to individual do not seem feasible in a much larger economy. An identical problem of distributing a surplus among many people arises if it is desired to improve income distribution.

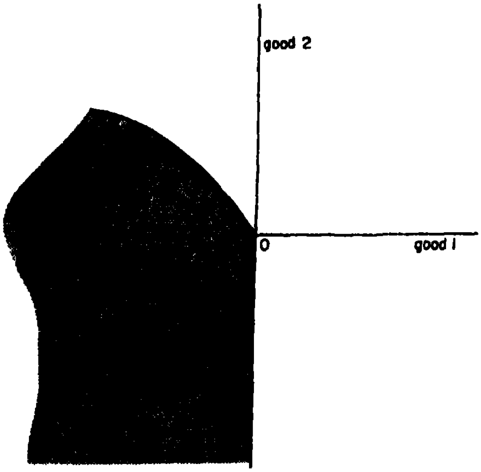


FIGURE 1

Thus we shall consider the use of commodity taxes when lump sum taxes are not permitted to the government, not for the intrinsic interest of this question in a one-consumer economy, but as an introduction to the many-consumer case. Furthermore we shall hold constant the government expenditure pattern which directly affects consumer utility. Thus we can ignore it, since the utility function already reflects its impact. The addition of choice for public consumption will be considered in Diamond-Mirrlees II.

Assuming free disposal, the technological constraint on the planner is that the government supply be on or under the production frontier. Such a constraint is shown by the shaded area in Figure 1. Let us measure on the axes the quantities supplied to the consumer. Thus, the output being produced (good 2) is measured positively, while the input (good 1) is measured negatively. The case drawn is the familiar one of decreasing returns to scale. If the government needed a fixed bundle of resources, for national defense say, then the production possibility frontier (describing the potential transactions with the consumer) would not pass through the origin. With constant returns to scale this might appear as in Figure 2, where a units of good 1 are needed for defense. (It is perhaps convenient to think of good 1 as labor and good 2 as a consumption good.)

In a totally planned economy, where the planner selects a fixed consumption bundle (including labor to be supplied) for each consumer, the planner would have no further constraint and could choose any point that was technologically feasible. Again, this is not implausible for the planner in a one-consumer economy, but becomes so as the number of households grows. A more realistic assumption, then, is to assume that the planner can only deal with consumers through the market place, hiring labor and selling the consumer good. Assume further that the planner is constrained to charge uniform prices. The planner must now set the price of the consumer good relative to the wage (or inversely the real wage), and is constrained to transactions which the consumer is willing to undertake at some relative price. The locus of consumption bundles which the consumer is willing to achieve by trade from the origin is the offer curve or price-consumption locus. It represents the

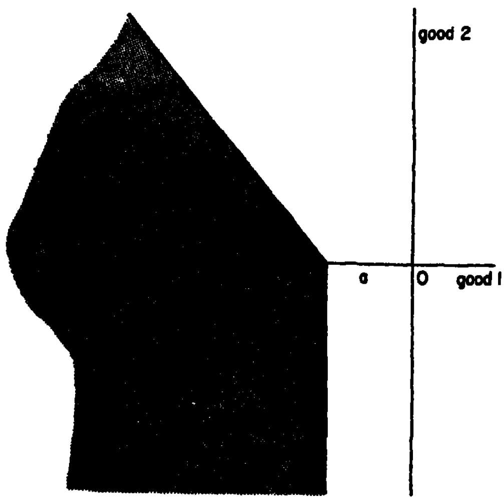


FIGURE 2

bundles of goods that the consumer would purchase at different possible price ratios. Figure 3 contains an example of an offer curve with several hypothetical budget lines and the corresponding indifference curves drawn in. The planner thus has two constraints: he must choose a point which is both technologically feasible and an equilibrium bundle from the point of view of the consumer. Combining these two constraints, the range of consumption bundles which are both feasible and potential consumer equilibria is shown as the heavy line in Figure 4.

We can state these two constraints algebraically. Let us denote by $z = (z_1, \ldots, z_n)$ the vector of government supply. The production constraint is then written

(1) $G(z) \leqq 0$, or, equivalently,

$$z_1 \leqq g(z_2, z_3, \ldots, z_n)$$

The constraint that the government supply equal the consumer demand for some price can be written in vector notation

(2) $$x(q) = z,$$

where $x = (x_1, \ldots, x_n)$ is the vector of consumer demands and $q = (q_1, \ldots, q_n)$ is the vector of prices faced by the consumer.

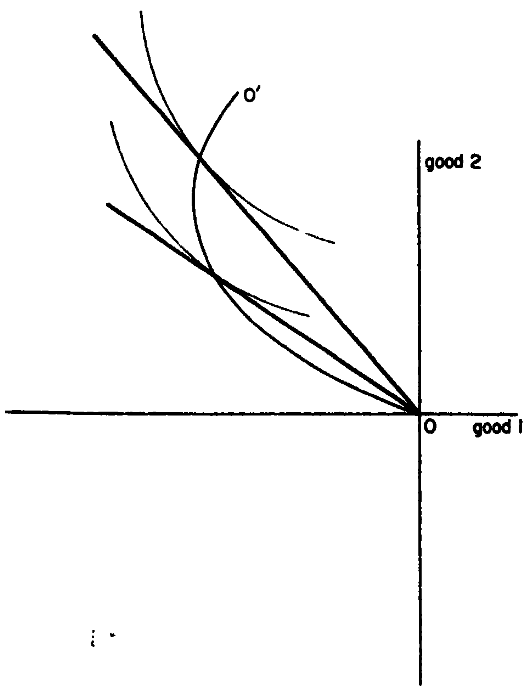


FIGURE 3

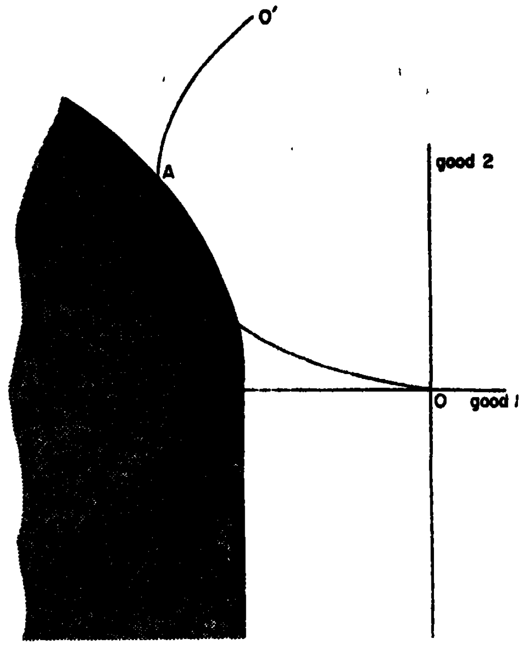


FIGURE 4

Now consider the government's objectives. Since the consumer's equilibrium position is determined by the prices he faces, we can, in the usual circumstances, describe the objective function as a function of prices, say $v(q)$. The problem is to choose q so as to

(3) Maximize $v(q)$

subject to $G(x(q)) \leqq 0$

This simply formulated problem is the focus of attention of the paper and can take on a variety of interpretations. The reader may note that the consideration of many consumers does not alter the form of this problem. This is a major advantage

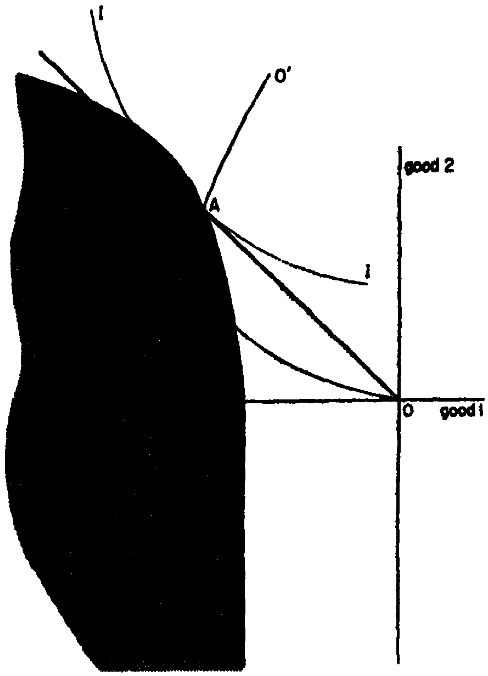


FIGURE 5

of using prices rather than quantities as the focus of the analysis.

Let us consider the case where the planner seeks to maximize the same function of consumption as the consumer's utility function. The welfare function is said to be *individualistic*, or to respect individual preferences, since welfare can be written as a function of individual utility. Returning to Figure 3 we see that the consumer moves to higher indifference curves as he proceeds along the offer curve away from the origin. Thus, in Figure 4 we wish to move as far along OO' as possible, subject to the constraint of the shaded production possibility set. The optimal point is therefore A, where the offer curve and the production frontier intersect.

The prices which will induce the consumer to purchase the optimal consumption bundle are defined by the budget line OA. In Figure 5 we show the optimal point and the implied budget line, and indifference curve II. All the points above II and in the shaded production set are Pareto-superior to A and technologically feasible, but not attainable by market transactions without lump sum transfers. For contrast, in Figure 6, we show the Pareto optimal point, B, and the implied budget line, and indifference curve $I'I'$, which will permit decentralization. In the case drawn, the consumer's budget line does not pass through the origin; this represents his payment of a lump sum tax to cover government expenditures in excess of profits from production.

We see that the optimal point is on the production possibility frontier of the economy, not inside it. This important property of the optimum can easily be seen to carry over to the case of many commodities, but still one consumer. With many commodities, the offer curve is a union of loci, each of which is obtained by holding the prices of all but one commodity constant and varying the price of that one commodity. Doing this for each com-

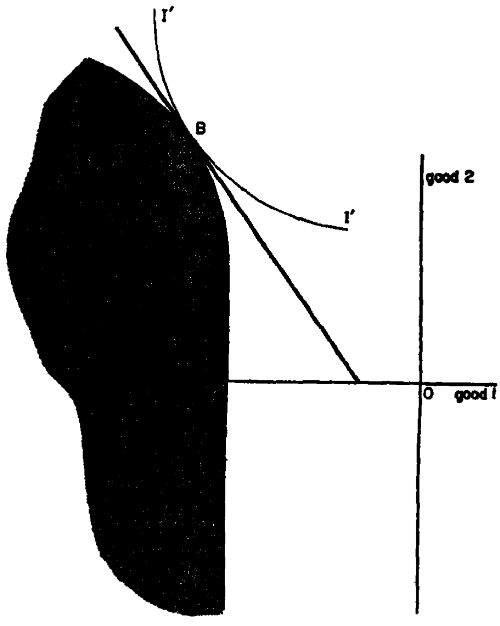


FIGURE 6

modity, and for all possible configurations of prices for the other commodities, generates all the loci. The offer curve is the union of such loci. On each locus, the point which is also on the production frontier is better than the other points on the locus. Thus, any point which is not on the production frontier is dominated by some point which is on the frontier. Therefore, the optimal point is one of the points on the frontier. The implications of this result will be seen more clearly below, when we consider both public and private production. For this result to carry over to the case of many consumers requires one further, mild assumption which will be discussed in the third section. First, we treat the one consumer economy algebraically, with both public and private production, showing by calculus the desirability of aggregate production efficiency, and obtaining the optimal relationship between consumer prices and the slope of the production possibilities. This relationship defines the optimal tax structure.

II. One-Consumer Economy—Algebraic Analysis

We assume constant returns to scale in the private production sector and the presence of competitive conditions there. In equilibrium there are, therefore, no profits. (This is a critical assumption for the efficiency analysis.) We also assume, for the present, that the only taxes used by the government are commodity taxes.[6] Consumer prices, q, therefore determine the choices available to the consumer, and we may write the welfare function as a function of consumer prices, $v(q)$. Notice that this covers the case where the government's assessment of welfare does not coincide with the consumer's utility, although depending on what he consumes. In the special case where social preferences coincide with those of the single consumer, his utility may be taken to measure welfare. Then we have

$$v(q) = u(x(q)) \tag{4}$$

We shall not use this special form for $v(q)$ in the analysis below until we come to evaluate the tax structure explicitly. Until that point, the analysis applies also to welfare functions that are not individualistic. For later use let us express the derivatives of v in this special case. Writing $v_k = \partial v/\partial q_k$, $u_i = \partial u/\partial x_i$, and using (4), we have

$$v_k = \sum u_i \frac{\partial x_i}{\partial q_k} = -\alpha x_k, \tag{5}$$

where α is a positive constant (i.e., independent of k), the marginal utility of income. Equation (5) follows from the budget constraint,

$$\sum q_i x_i = 0, \tag{6}$$

which on differentiation with respect to q_k yields

$$x_k + \sum q_i \frac{\partial x_i}{\partial q_k} = 0 \tag{7}$$

Since utility-maximization implies that $u_i = \alpha q_i$, (5) now follows from (7).

Production

Let us denote the vector of prices faced by private producers by $p = (p_1, \ldots, p_n)$. Because of taxes, t, these may differ from the prices faced by consumers: $q_i = p_i + t_i$ $(i = 1, \ldots, n)$. $y = (y_1, \ldots, y_n)$ is the vector of commodities privately supplied (inputs will thus appear as negative supplies), and we write the private production constraint,

$$y_1 = f(y_2, \ldots, y_n) \tag{8}$$

Notice that we assume *equality* in the

[6] This assumption is made solely for simplicity. In Diamond-Mirrlees II, the general principles will be seen to carry over with additional taxes, including a progressive income tax.

production constraint, that is, that production is efficient in the private sector. This follows from profit maximization if there are no zero prices. We assume that f is a differentiable function, and that $y_i \neq 0$ $(i=1, \ldots, n)$. Then, profit maximization means that

$$p_i = -p_1 f_i(y_2, \ldots, y_n), \quad (i = 2, \ldots, n) \tag{9}$$

where f_i denotes the derivative of f with respect to y_i. Also, by the assumption of constant returns to scale, maximized profits are zero in equilibrium:

$$\sum p_i y_i = 0 \tag{10}$$

So that we may conveniently employ calculus, we shall assume that the government production constraint, (1), is satisfied with an equality rather than an inequality:

$$z_1 = g(z_2, \ldots, z_n) \tag{11}$$

Thus we do not give the government the option of inefficient government production. Rather, we shift our attention to *aggregate* production efficiency. Efficiency will be present if marginal rates of transformation are the same in publicly and privately controlled production. It will then be seen quite easily that the assumption of efficiency in the public sector is justified.

Walras' Law

We have chosen an objective function and expressed the government's production constraint above. To complete the formulation of the maximization problem, it remains to add the requirement that the economy be in equilibrium. The conditions that all markets clear can be stated in terms of the vectors x, y, and z.

$$x(q) = y + z \tag{12}$$

The reader may be puzzled that at no place in this formulation has a budget constraint been introduced for the government. (Other readers may be puzzled by our failure to include only $n-1$ markets in our market clearance equations. These are aspects of the same phenomenon.) Walras' Law implies that if all economic agents satisfy their budget constraints and all markets but one are in equilibrium, then the last market is also in equilibrium. It also implies that when all markets clear and all economic agents but one are on their budget constraints, then the last economic agent is on his budget constraint. In setting up our problem, we have assumed that the household and the private firms are on their budget constraints. Thus, if we assume that all markets clear, this will imply that the government is satisfying its budget constraint,[7] which we can express as

$$\sum (q_i - p_i)x_i + \sum p_i z_i = 0 = \sum t_i x_i + \sum p_i z_i \tag{13}$$

Alternatively, if we consider the government budget balance as one of the constraints, then it is only necessary to impose market clearance in $n-1$ of the markets.

In this model we can make two price normalizations, one for each price structure. Since both consumer demand and firm supply are homogeneous of degree zero in their respective prices, changing either price level without altering relative prices leaves the equilibrium unchanged. As normalizations let us assume,

$$p_1 = 1, \qquad q_1 = 1, \qquad t_1 = 0 \tag{14}$$

It may seem surprising that it does not matter whether the government can tax good one. But the reader should remember the budget balance of the consumer. Since there are no lump sum transfers to the

[7] In an intertemporal interpretation of this model, the government budget is in balance over the horizon of the model, not year by year.

consumer, net expenditures are zero. Thus, levying a tax at a fixed proportional rate on all consumer transactions results in no revenue. (It should be noticed that a positive tax rate applied to a good supplied by the consumer is in effect a subsidy and results in a loss of revenue to the government.)

Welfare Maximization

We can now state the maximization problem. In the statement we shall use the two sets of prices as control variables. It would be a more natural approach to use the taxes which the government actually controls as decision variables. However, once we have determined the optimal p and q vectors we have determined the optimal taxes. Using taxes as decision variables complicates the mathematical formulation and leads to a control problem since the tax vector may not uniquely determine equilibrium.

Rather than calculate the first-order conditions from the formulation spelled out above, we shall alter the problem to simplify the derivation. We have to choose

$$(15)\quad q_2, \ldots, q_n, \quad p_2, \ldots, p_n, \quad z_1, \ldots, z_n$$

to maximize $v(q)$ subject to

$$x_i(q) - y_i - z_i = 0 \qquad (i = 1, 2, \ldots, n),$$

where y maximizes $\sum p_i y_i$ subject to

$$y_1 = f(y_2, \ldots, y_n),$$

and

$$z_1 = g(z_2, \ldots, z_n)$$

Since the choice of producer prices can be used to obtain any desired behavior on the part of private producers, we can use any vector y consistent with the production constraint (8). Producer prices are then determined by equation (9). Using the equations

$$y_2 = x_2 - z_2, \ldots, y_n = x_n - z_n,$$

we reduce the constraints in (15) to the single constraint

$$x_1(q) = y_1 + z_1$$
$$= f(x_2 - z_2, \ldots, x_n - z_n) + g(z_2, \ldots, z_n)$$

We have therefore simplified the problem (15) to:

$$(16)\quad \text{Choose } q_2, \ldots, q_n, \qquad z_2, \ldots, z_n$$

to maximize $v(q)$ subject to

$$x_1(q) - f(x_2(q) - z_2, \ldots, x_n(q) - z_n) - g(z_2, \ldots, z_n) = 0$$

Forming a Lagrangian expression from (16), with multiplier λ,

$$(17)\quad L = v(q) - \lambda[x_1(q) - f(x_2 - z_2, \ldots, x_n - z_n) - g(z_2, \ldots, z_n)],$$

we can differentiate with respect to q_k:

$$(18)\quad v_k - \lambda\left(\frac{\partial x_1}{\partial q_k} - \sum_{i=2}^{n} f_i \frac{\partial x_i}{\partial q_k}\right) = 0 \qquad k = 2, 3, \ldots, n$$

Making use of the equations (9) for producer prices, this can be written

$$(19)\quad v_k - \lambda \sum_{i=1}^{n} p_i \frac{\partial x_i}{\partial q_k} = 0 \qquad k = 2, 3, \ldots, n$$

Differentiating L with respect to z_k we have

$$(20)\quad \lambda(f_k - g_k) = 0 \qquad k = 2, 3, \ldots, n$$

Provided that λ is unequal to zero (i.e., that there is a social cost to a marginal need for additional resources), equation (20) implies equal marginal rates of transformation in public and private production and thus aggregate production efficiency as was argued above. The assumption that $\lambda \neq 0$ needs justification. This is provided by the rigorous arguments of Sections III and IV.

If we had introduced *several* public

production sectors, each described by a constraint like (11), we should have obtained an equation of the form (20) for *each* sector. Thus marginal rates of transformation in all public sectors should be equal, since they are all to be equal to the private marginal rates of transformation. This argument—which we only sketch here, since the conclusion will be proved more directly in the next section—justifies our assumption that there should be production efficiency in the public sector.

Optimal Tax Structure

The relations (19) determine the optimal tax structure, since they show how producer and consumer prices should be related. These equations show that consumer prices should be at a level such that further increases in any price result in an increase in social welfare, v_k, which is the same ratio, λ, to the cost of satisfying the change in demand arising from the price increase. Reintroducing taxes explicitly into the problem we can obtain an alternative interpretation for the first-order conditions.

Since x_i is a function of $p+t$,

$$\frac{\partial x_i}{\partial q_k} = \frac{\partial x_i}{\partial t_k}$$

(p is held constant in this latter derivative.) Consequently, the optimal tax structure, (19), can be rewritten:

$$(21) \qquad v_k = \lambda \sum p_i \frac{\partial x_i}{\partial t_k} = \lambda \frac{\partial}{\partial t_k} \sum p_i x_i$$

Since $\sum p_i x_i = \sum q_i x_i - \sum t_i x_i = -\sum t_i x_i$ (by the consumer's budget constraint (6)), we have

$$(22) \qquad v_k = -\lambda \frac{\partial}{\partial t_k} \left(\sum t_i x_i\right)$$

This last set of equations asserts the proportionality of the marginal utility of a change in the price of a commodity to the change in tax revenue resulting from a change in the corresponding tax rate, calculated at constant producer prices. Like the first-order conditions for the optimum in standard welfare economics, our first-order conditions are expressions in constant prices. The tax administrator, like the production planner, need not be concerned with the response of prices to government action when looking at the first-order conditions.

If we now make the further assumption that the welfare function is individualistic, we can use equation (5) to replace v_k. The first-order conditions then become

$$(23) \qquad x_k = \frac{\lambda}{\alpha} \frac{\partial\left(\sum t_i x_i\right)}{\partial t_k}$$

Thus for all commodities the ratio of marginal tax revenue from an increase in the tax on that commodity to the quantity of the commodity is a constant. This form of the first-order conditions has the advantage of showing the information needed to test whether a tax structure is optimal. The amount of information does not seem excessive relative to the data and knowledge which a planner in an advanced country should have.

The statements of the first-order conditions thus far do not directly indicate the size of the tax rates required, nor the impact upon demand that the optimal tax rates would have. In his pioneering study of optimal tax structure, Frank Ramsey manipulated the first-order conditions so as to shed light on the latter question. He employed the concept of demand curves calculated at a constant marginal utility of income. Paul Samuelson reformulated this using the more familiar demand curves calculated at a constant level of utility. We shall return to this question in Diamond-Mirrlees II.

III. Production Efficiency in the Many-Consumer Economy

We have remarked already that many of the results carry over directly to an economy of many consumers, even when lump sum taxation is excluded. We notice at once that the device of expressing welfare as a function of the prices, q, faced by consumers can be used perfectly well. Explicitly, we assume that there are H households, with utility and demand functions u^h and x^h $(h=1, 2, \ldots, H)$. If, as we may generally suppose, in the absence of externalities from producers to consumers, social welfare can be expressed as a function of the consumption of the various consumers in the economy, $U(x^1, x^2, \ldots, x^H)$, it may also be written

$$(24) \quad V(q) = U(x^1(q), x^2(q), \ldots, x^H(q)),$$

where we assume that there are no lump sum incomes or transfers that would be influenced by producer prices or government policy. In the case where social welfare depends only on individual utility and there are no externalities, we can write

$$(25) \quad V(q) = W[u^1(x^1(q)), u^2(x^2(q)), \ldots, u^H(x^H(q))],$$

where W is presumed to be strictly increasing in each of its arguments.

Using this indirect welfare function, we can carry out the analysis already presented for the one-consumer economy, and conclude in the same way that aggregate production efficiency is desirable. For that argument to be correct, we must confirm that the Lagrange multiplier λ is not zero. Rather than attempt to do this directly, we shall present a different argument for the desirability of production efficiency. A further condition will be required to secure our conclusion. In considering this problem, we shall concentrate on the case where all production is under government control. The desirability of production efficiency in this case will be seen to imply the same conclusion when there is also a private sector (provided that private producers are price takers, and profits, if any, are transferred to the government). Assume then (as we did in Section I) that all production takes place in the public sector: our problem is to find q that will

$$(26) \quad \text{Maximize } V(q), \qquad \text{subject to } G(X(q)) \leq 0,$$

where we define $X(q) = \sum_h x^h(q)$ as aggregate demand at prices q. We shall also express the production constraint a little more generally by saying that $X(q)$ is to belong to the production set G, the set of technologically feasible production plans. (Thus the letter G denotes both the production set, and also the function that can be used to describe it; but we shall hardly ever use the *function* G explicitly).

Suppose we establish that, at the optimum for problem (26), production is efficient. Consider an economy with the same technolgical possibilities, partly under the control of private, competitive producers. The government can induce private firms to produce any efficient net output bundle by suitable choice of producer prices p. In particular, it can obtain the production plan that would be optimal if the government controlled all production. The choice of p does not affect consumer demands or welfare, since pure profit arising from decreasing returns to scale go to the government, and since, any commodity taxes being possible, q can be chosen independently of p. Thus, if the solution to (26) is efficient, the same equilibrium can be achieved when some production is under private control, and is optimal in that case too. Proof that production efficiency is desirable in the "special" case (26) therefore implies that pro-

duction efficiency is desirable in the more general case.

Examples of Inefficiency

Before considering the argument for efficiency, it is useful to consider some limitations on that argument as demonstrated by the following examples of desired inefficiency. It will be recollected that a production plan is efficient if any other feasible production plan provides a smaller net supply of at least one commodity. We shall use a different concept: we say that a production plan is *weakly efficient* if it is on the production frontier. It is possible for a production plan to be weakly efficient without being efficient if the production frontier has vertical or horizontal portions. For matters of economic importance, such as the existence of shadow prices, weak efficiency is all that is required. It is easy to see that if all the prices corresponding to a weakly efficient production plan are positive, the plan is in fact efficient in the usual sense.

Even with this slightly weakened concept of efficiency, it is not necessarily true that, when an optimum exists, optimal production has to be weakly efficient. We present two examples.

Example a is portrayed in Figure 7. It is a one-consumer economy where social preferences, as depicted in the social indifference curve *II*, do not coincide with individual preferences. It is evident that, in the case shown, the optimal production plan is actually in the interior of the production set.

In the second example, social preferences do respect household preferences, but again optimal production lies in the interior of the production set, and is therefore not weakly efficient: suitable producer prices cannot be found, and the social optimum cannot be obtained when there is private control of production.

Example b. There are two commodities and two households. One has utility function x^2y, the other has utility function xy^2; each has the nonnegative quadrant $\{(x, y) | x \geq 0,\ y \geq 0\}$ as consumption set. The first consumer has three units of the first commodity initially; the second, one unit of the second commodity. The welfare function is

$$-\frac{1}{x_1y_1}-\frac{1}{x_2y_2}$$

The second commodity can be transformed into the first according to the production relation $x+10y\leq 0,\ (x\geq 0)$. Let the prices of the commodities be q_1, q_2. Then the first household's net demands are

$$-1 \text{ of the first commodity,}$$

$$q_1/q_2 \text{ of the second commodity.}$$

The second household has net demands

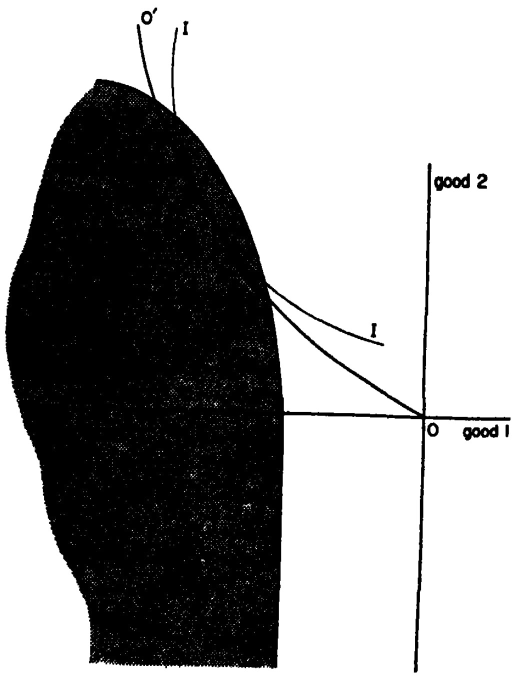


FIGURE 7

$$\tfrac{1}{3}(q_2/q_1) \quad \text{and} \quad -\tfrac{1}{3}$$

Thus, the net market demand for the commodities is

$$x = \tfrac{1}{3}(q_2/q_1) - 1 \quad \text{and} \quad y = (q_1/q_2) - \tfrac{1}{3}$$

These must satisfy

$$x + 10y \leqq 0, \qquad x \geqq 0$$

Welfare is $-q_2/4q_1 - 27q_1/4q_2$ which is maximized when $q_2/q_1 = 3\sqrt{3}$: the corresponding production vector $\sqrt{3}-1, \tfrac{1}{3}(\sqrt{\tfrac{1}{3}} - 1)$ is actually interior to the production set, not on the frontier. This example has the unimportant peculiarity that initial endowments of the consumers are on the frontiers of their consumption sets. More complicated examples avoiding the peculiarity have been constructed.

The Efficiency Argument

Despite these examples, the following argument shows that optimal production will generally be on the production frontier. Suppose that the aggregate demand functions, $X(q)$, are continuous. Then any small change in the prices, q, will not change aggregate production requirements by much. Therefore, if optimal production were in the interior of the production set, small changes in consumer prices would still result in technologically feasible aggregate demands. Thus, if we are at the optimum, small changes in consumer prices cannot increase welfare. If we can argue that, at the optimum, there exists a small price change which would increase $V(q)$, we can conclude that production for the optimum must occur on the production frontier. For any unsatiated single consumer, utility can be increased either by lowering the price of a supplied good or raising the price of a demanded good (as we can see, algebraically, in equation (5)). With a single consumer, we need not argue further, provided the equilibrium involves some trade. When there are many consumers, we can be certain of increasing welfare if we raise some consumer's utility without lowering that of anyone else. If there is a commodity that no consumer purchases, but some consumer supplies (such as certain labour skills); or a good (with positive price) which no consumer supplies, but some consumer purchases (such as electricity), we could alter the price of that commodity in such a way as to bring about an unambiguous increase in welfare. In that case, we conclude that efficient production is required for the maximization of individualistic social welfare. In example *b*, it will be seen that neither of the commodities is supplied, or demanded, by both consumers. The very simplicity of the case appears to be misleading.

A formal presentation of this argument is given in the next section: these technical details can be omitted without loss of continuity. We conclude this section by introducing further taxes into the discussion.

First, consider the case of a poll tax (or subsidy)—that is, a tax is paid by a household on the basis of some unalterable property, such as its sex or age distribution. Such a tax is, of course, a lump sum tax, although its availability is not, in general, sufficient to enable the full optimum to be achieved. To fix ideas, suppose there is a single transfer, τ, to be made to all households. Then welfare can be written $V(q, \tau)$, and we are to

$$\text{(27)} \quad \text{Maximize } V(q, \tau) \quad \text{subject to } X(q, \tau) \text{ being in } G$$

The standard efficiency argument can be used. Let (q^*, τ^*) be the optimum: if any small change in q or τ would increase V, optimal production, $X(q^*, \tau^*)$ must be on the production frontier (assuming that X is a continuous function). Now a poll subsidy must make everyone better off,

unless some are already satiated, and so must a small increase in subsidy. Thus so long as a poll subsidy is possible (and it surely is) and not every household is satiated, optimal production must be on the frontier.

Adding further tax instruments to the government's armoury in no way weakens the efficiency conclusion. We simply note that if there are other tax variables which are independent of producer prices and quantities, denoted collectively by ζ, we can hold them constant at their optimum values ζ^*, and then apply the efficiency argument to the problem (27) or (26), where V and X are evaluated for $\zeta=\zeta^*$.

Our final conclusion is that whatever the class of possible tax systems, if all possible commodity taxes are available to the government, then in general, and certainly if a poll subsidy is possible, optimal production is weakly efficient. We would not expect this conclusion to be valid if there were constraints on the possibilities of commodity taxation, or more generally, on the possible relationship between producer prices and consumer demand. The presence of pure profits is one example of such a relationship. To show what goes wrong, suppose, by way of another example, that *no* commodity taxes are possible, but a poll tax is possible, and that part of production is privately controlled, in such a way that it is uniquely determined by producer prices. Then we have to choose a public production vector z and a poll tax τ to

(28) Maximize $V(p, \tau)$

subject to $X(p, \tau) - y(p) = z$ being in G,

where $y(p)$ is the private production vector when prices are p. Following the argument used above, we consider τ smaller than τ^*, the optimum level, and note that $V(p^*, \tau) > V(p^*, \tau^*)$. This implies that $X(p^*, \tau) - y(p^*)$ is not in G, and therefore z^*, the optimal z, is efficient in G. But the argument does not imply that the aggregate optimal production plan, $y(p^*)+z^*$ is efficient. Of course, in an economy where all production is under public control, these problems do not arise. Even when some of the q_k are fixed, the efficiency argument holds, for there can be no necessary relation between q and p.

IV. Theorems on Optimal Production

In this section, we explore the existence of the optimum, and the efficiency of optimal production, rigorously. We rely on Debreu (1959) for the results of general equilibrium theory that are required.

Assumptions

There are H households in the economy, each household choosing a preferred net consumption vector x from his consumption set C subject to the budget constraint $q \cdot x \leqq 0$ where q is the vector of prices charged to consumers. (Consumption is measured net of initial endowment for convenience, since the latter is unaltered in the analysis.) As usual the net demand vector x has, in general, both positive and negative components corresponding to purchases and sales by the household.

The assumptions used below will be selected from the following list (the superscript h refers to the index of households; all assumptions, when made, hold for all h):

(a.1) C^h is closed, convex, bounded below by a vector a^h, and contains a vector with every component negative.

(a.2) The preference ordering is continuous.

(a.3) The preference ordering is strongly convex. Formally, if x^2 is preferred or indifferent to x^1 and $0<t<1$, then $tx^2+(1-t)x^1$ is strictly preferred to x^1.

(a.4) There is no satiation consumption in C^h.

Assumptions (a.1) and (a.2) guarantee the existence of continuous utility functions, which we shall write u^h (see Debreu Section 4.6). Furthermore, under (a.1)–(a.3), when the demand vector $x^h(q)$ is defined, it is uniquely defined. When C^h is bounded, assumptions (a.1)–(a.3) imply that $x^h(q)$ is defined and continuous at all non-zero nonnegative q. (See Debreu, Section 4.10.)

Let us denote aggregate demand by $X(q)=\sum_h x^h(q)$.

It is assumed that all production is controlled by the government. The assumptions on the production possibility set, G, will be taken from the following set:

(b.1) Every production plan in which nothing is produced in a positive quantity is possible: i.e., if $z\leqq 0$, z is in G.
(b.2) Complete inactivity is possible: i.e., 0 is in G.
(b.3) G is closed.
(b.4) There exists a vector $\bar{a}$ such that $z\leqq\bar{a}$ for all nonnegative z in the convex closure of G. (i.e., the closure of the convex hull of G).[8]
(b.5) G is convex.

The welfare function will be denoted by $U(x^1, \ldots, x^H)$. When demands are functions of prices only we can define the indirect welfare function as

$$V(q) = U(x^1(q), \ldots, x^H(q))$$

Similarly we can define an individual's indirect utility function by

$$v^h(q) = u^h(x^h(q))$$

We shall say that the welfare function *respects household preferences* when U can be written

$$U(x^1, \ldots, x^H) = W(u^1(x^1), \ldots, u^H(x^H))$$

[8] When G is convex, this assumption is similar to the assumption that inputs are required to obtain outputs, but permits the government to own a vector of inputs.

with W increasing in each argument. We shall assume

(c.1) U is a continuous function of $(x^1, \ldots, x^H)$

We can now state our problem as trying to find q^* to maximize $V(q)$ subject to $X(q)$ being in G. A commodity vector will be called *attainable* if it is feasible and if there exists prices such that aggregate demand equals the vector. The set of all such vectors, the *attainable set*, is the intersection of G with the set of vectors $X(q)$ for all nonnegative q.

Existence of an Optimum

If we assume that the attainable set is nonempty and bounded, we obtain

THEOREM 1. *If assumptions* (a.1)–(a.3), (b.3), *and* (c.1) *hold, and if the attainable set is nonempty and bounded, an optimum exists.*

PROOF:

Consider an economy in which the consumption sets are truncated by removing from them all points x with $\|x\|>M$, where all vectors in the attainable set satisfy $\|x\|<M$. For this truncated economy, the demand functions are continuous at all price vectors not equal to zero. Since the attainable set, and demands for any q corresponding to an attainable vector, are the same in the original and truncated economies, an optimum for the truncated economy is an optimum for the original economy. In other words, we may, without loss of generality, assume that demands are continuous at $q\neq 0$. Since the demand functions are homogeneous of degree zero in the prices, we can restrict our attention to q satisfying $q\geqq 0$ and $\sum_i q_i=1$.

We next demonstrate that the set $\{q\,|\,X(q) \text{ in } G\}$ is closed. Let q_n be a sequence of price vectors converging to q', with $X(q_n)$ in G for all n. Let x' be a limit point of $\{X(q_n)\}$. Since G is closed, x' is

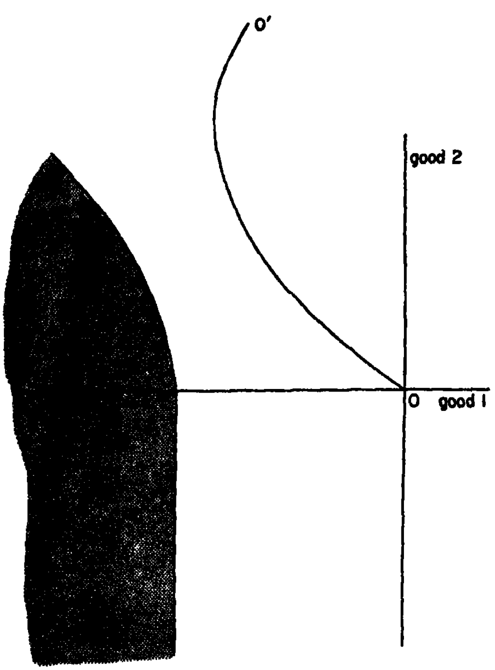


FIGURE 8

in G. At the same time, $x'=X(q')$, by the continuity of X. Thus q' is in $\{q|X(q) \text{ in } G\}$, which is therefore closed.

Since the attainable set is nonempty, and prices are in any case bounded, $\{q|X(q) \text{ in } G\}$ is closed, bounded, and nonempty. By the continuity of the demand functions, and assumption (c.1), V is a continuous function of q, which therefore attains its maximum on the set $\{q|X(q) \text{ in } G\}$.

One criterion for the attainable set to be nonempty follows immediately from the existence of competitive equilibrium in an exchange economy:

THEOREM 2. *If assumptions* (a.1)–(a.4) *and* (b.1) *hold, the attainable set is nonempty.*

PROOF:

See Debreu (Section 5.7) for a proof that there exists an equilibrium for the exchange economy with these consumers. The equilibrium prices result in a feasible demand.

If the production set is taken to be the set of possible production vectors net of government consumption, the assumption that zero production is possible is excessively strong, especially for governments with large military establishments. But it is easy to construct examples of economies not satisfying (b.1) in which there is no attainable point. Consider the one-consumer economy depicted in example c shown in Figure 8.

The boundedness of the attainable set would be implied by the boundedness of the consumption sets, or the boundedness of production, but the following case is more appealing:

THEOREM 3. *If assumptions* (a.1) *and* (b.2)–(b.4) *hold, then the attainable set is bounded.*[9]

PROOF:

Suppose the attainable set is not bounded. Then there exists a sequence of attainable vectors x_n such that $||x_n||$ is an unbounded increasing sequence of real numbers. There exists an n' such that $||x_{n'}||>||\bar{a}||$, where $\bar{a}$ is the vector employed in (b.4). Consider the sequence of vectors $(||x_{n'}||/||x_n||)x_n$ for $n\geq n'$. Each vector is in the convex hull of G (being a convex combination of the origin and x_n). Further the sequence is bounded. Thus there is a limit point, ξ, which is in the convex closure of G and satisfies $||\xi||>||\bar{a}||$. Let $b=\sum_h a_h$, where a_h are the vectors employed in (a.1). Then $x_n=\sum_h x_n^h \geqq \sum_h a_h=b$. Further $(||x_{n'}||/||x_n||)x_n \geqq (||x_{n'}||/||x_n||)b$. But the latter sequence of vectors converges to zero. Thus $\xi\geqq 0$. This is a contradiction.

[9] The attainable set will also be bounded if (b.2)–(b.4) hold for the true production set, *gross* of government consumption, rather than the net production set, G. Thus the assumption that zero production is possible is not of great consequence.

Finally, we should remark that the strong convexity assumption, (a.3), which was made in Theorem 1 can be changed to convexity without affecting the conclusion. All that is required is to replace the continuous functions of the proof by upper semi-continuous correspondences. On the other hand, one can easily construct examples in which an optimum fails to exist because of the absence of continuity.

Efficiency

The following lemma provides two criteria for optimal production to be on the frontier of the production set. It will be used to deduce a theorem about the case where household preferences are respected.

LEMMA 1: *Assume an optimum, q^*, exists. If aggregate demand functions and the indirect welfare function are continuous in the neighborhood of the optimal prices; and if either*

(1) *for some i, V is a strictly increasing function of q_i in the neighborhood of q^*; or*

(2) *for some i with $q^*>0$, V is a strictly decreasing function of q_i in the neighborhood of q^*,*

then $X(q^*)$ is on the frontier of G.

PROOF:

Let l_i be the vector with all zero components except the ith, which is one. In case 1, for ϵ sufficiently small $V(q^*+\epsilon l_i) > V(q^*)$. Hence $X(q^*+\epsilon l_i)$ is not in G. Letting ϵ decrease to zero, the continuity of X shows that $X(q^*)$ is a limit of points not in G, and therefore belongs to the boundary of G. In case 2, a similar argument can be made using $V(q^*-\epsilon l_i)$.

These conditions are weak. They are, naturally, independent of production possibilities. It may also be noticed that, when V is a differentiable function of prices, the stated conditions are equivalent to assuming that

(29) It is not the case that $V'(q^*) \leqq 0$

Here $V'(q)$ is the vector of first derivatives of V with respect to prices. The equivalence of the conditions of the theorem and (29) is clear if we remember that

$$V'(q)\cdot q = \sum \frac{\partial V}{\partial q_k} q_k = 0, \tag{30}$$

since V is homogeneous of degree zero in q. Therefore $V' \leqq 0$ if, and only if, $\partial V/\partial q_k = 0$ when $q_k > 0$ and $\partial V/\partial q_k \leqq 0$ in any case.

In the following theorem, we strengthen the assumptions in a different way: they remain notably weak.

THEOREM 4. *If* (a.1)–(a.4) *and* (c.1) *hold; if social welfare respects individual preferences; and if either*

(1) *for some i, $x_i^h \leqq 0$ for all h, and $x_i^{h'} < 0$ for some h'; or*

(2) *for some i, with $q_i > 0$, $x_i^h \geqq 0$ for all h and $x_i^{h'} > 0$ for some h';*

Then if an optimum exists, production for the optimum is on the frontier of the feasible set.

PROOF:

Individual demand functions are continuous in the neighborhood of the optimum and thus aggregate demands and the indirect welfare function are continuous. Since social welfare respects preferences, indirect social welfare can be written as an increasing function of indirect utilities. In case 1, indirect utilities are a nondecreasing function of q_i in the neighborhood of q^* for all h while the indirect utility function of h' is strictly increasing in q_i. Thus V increases with q_i. Case 2 follows similarly.

The assumption of strictly convex preferences made in Theorem 4 is required in the theorem as stated.

Example d: Consider an economy with one consumer whose indifference curves have

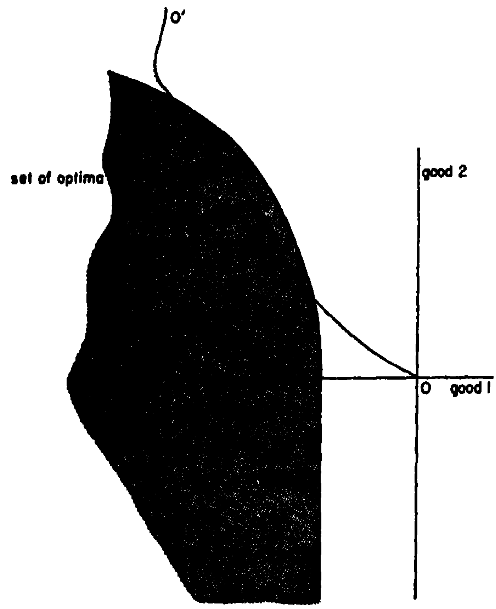


FIGURE 9

a linear section. Then the offer curve may coincide with the linear part of an indifference curve, giving a set of optima, only one of which is on the production frontier. As an illustration, see Figure 9.

The example suggests that we weaken the conclusion of Theorem 4 to say that there exists an optimum on the frontier of G: this generalization is indeed correct if we merely assume convexity of preferences. The proof follows that of Theorem 4, with upper semi-continuity of the demand correspondence replacing continuity of demand functions.

V. Extensions

We can summarize the efficiency result by considering an economy with three sectors—consumers, private producers, public producers. We assumed that only the equilibrium position of the consumer sector enters the welfare function, and that only market transactions take place between sectors, while the government has power to tax any intersector transaction at any desired rate. One conclusion was that all sectors not containing consumers should be viewed as a single sector, and treated so that aggregate production efficiency is achieved. By regrouping the parts of the economy according to this schematic division, we can extend the efficiency result to several other problems. In each case, we indicate briefly how application of this schematic view shows the relationship of the extension to the basic model.

Intermediate Good Taxation

The model, as presented above, left no scope for intermediate good taxation. If we separate private production possibilities into two (or many) sectors, we introduce the possibility of taxing transactions between firms. In the schematic view presented above, we could consider a consumer sector and two, constant returns to scale, private production sectors. We conclude that we want efficiency for these private production possibilities taken together. Therefore the optimal tax structure includes no intermediate good taxes, since these would prevent efficiency. (Similarly we conclude that government sales to firms should be untaxed while those to consumers are taxed.)

There is a straightforward interpretation of this result, which helps to explain the desirability of production efficiency. In the absence of profits, taxation of intermediate goods must be reflected in changes in final good prices. Therefore, the revenue could have been collected by final good taxation, causing no greater change in final good prices and avoiding production inefficiency. This interpretation highlights the necessity of our assumption of constant returns to scale in privately controlled production.

However, it may well be desirable to tax transactions between consumers or to charge different taxes on producer sales to

different consumers. There are two ways in which we can consider doing this. The country might be geographically partitioned with different consumer prices in different regions. Ignoring migration, and consumers making purchases in neighboring regions, our analysis can be applied to determine taxes region by region. In general the tax structure will vary over the country.

Alternatively, we might consider taxation on all consumer-consumer transactions. Here, too, we would expect to be able to increase social welfare by having these additional tax controls. Neither addition to the available tax structure alters the desirability of production efficiency.

Untaxable Sectors

One problem that arises with a model considering taxation of all transactions is that some transactions may not be taxable, practically or legally. An example of the former might be subsistence agriculture where transactions with consumers are hard to tax while those with firms are not. If the introduction of other taxes (e.g., on land or output) is ruled out, we can accommodate this problem in the model by including subsistence agriculture in the consumer rather than producer sector (or treating it as a second consumer sector). Efficiency would then be desired for the modern and government production sectors taken together; while the tax structure rules would be stated in terms of demand derivatives of the augmented consumer sector rather than of just the true consumers.

Similarly, in an economy without taxes, a public producer subject to a budget constraint is unable to charge different prices to consumers and producers. Lumping together the entire private sector as a single consumer sector, we obtain the conditions for optimal public production of an industry regulated in this manner. This is the problem considered by Boiteux in the context of costless income redistribution. He also analyzed such an economy with several firms, each limited by a budget constraint.

Foreigners

It is not easy to provide a satisfactory welfare economics for a world of many countries. The study of world welfare maximization is interesting, and, one may hope, "relevant." But it has the serious limitation that its results can seldom be applied to the actions of governments. However altruistic the principles on which a government seeks to act, it has to allow for the actions other governments may take, based on different principles, or for different reasons. (A somewhat analogous problem arises in intertemporal welfare economics.) In the following two subsections, we shall, in order to keep the discussion brief, refer only to the case where the reactions of all other countries are well-defined functions of the actions of the country directly considered. Thus we neglect, reluctantly, those situations that have come to be called "game-theoretic." Also, we shall not consider the problem of formulating a social welfare function in an international setting.

International Trade

So long as we are completely indifferent to the welfare of the rest of the world, and so long as the reactions of other countries are well-defined, international trade simply provides us with additional possibilities for transforming some goods and services into others. The efficiency result then implies that we would want to equate marginal rates of transformation between producing and importing. If there is a monopoly position to be exploited, it should be. If international prices are unaffected by this country's demand, intermediate goods should not be subject to a tariff, but final

good sales direct to consumers should be *subject to a tariff equal to the tax on the* same sale by a domestic producer.

Sometimes it is not possible to sell goods to foreigners at prices different from those at which they are sold to domestic consumers, although the theory just outlined suggests that foreigners should be treated like producers. As examples, we may cite tourists and commodities covered by special kinds of international agreement. If tourism, say, is an important trading opportunity for the country, and tourists have to be charged the same prices as domestic consumers, this will affect the optimal level of taxes on certain commodities. The general efficiency result is not upset, however. The analysis can be performed by treating tourists as consumers whose income does not affect social welfare.

The authors do not, of course, recommend indifference to the welfare of the rest of the world; although it happens to make the results somewhat neater. International trade provides the country with another set of consumers who can trade with it at prices different from its own consumers: the case (when foreign reactions are well-defined) is similar to the possibility of using different consumer prices in different regions of the same economy. In that case, there is no reason why optimal international trade prices should be the same as producer prices, p, or domestic consumer prices, q.

Migration

In all that has gone before, we have been holding constant the set of consumers in the economy. We can introduce migration in a straightforward manner. Social welfare may be a function of the consumption of every household in the world. Changes in the consumer prices charged in the home country cause migration in one direction or another, and therefore affect welfare in ways we have not previously discussed (such as the effect on the inhabitants of another country of having additional taxpayers join them). But we can still define an indirect welfare function $V(q)$, so long as the reactions of the rest of the world are well-defined. Similarly we can define aggregate demand functions $X(q)$, but these are no longer continuous. For, when a man decides to emigrate, his contribution to aggregate demand changes from x^h to 0.[10] But the number of migrants arising from a small price change may, quite reasonably, be assumed small relative to the population as a whole. We can therefore adequately approximate this situation by considering a continuum of consumers. In this way we can restore continuity to aggregate demand, and to the indirect welfare function. It is to be expected, then, that production efficiency is still desired. Since the derivatives of the demand functions, and possibly also the derivatives of V, will be different when the possibility of migration is allowed for, the optimal tax structure will be changed to reflect the loss of tax revenue when net taxpayers, for example, leave the country. While we do not wish to examine this problem in detail here, we believe that these ideas provide an interesting approach to the analysis.

Consumption Externalities

The schematic view of this problem given above suggests that the basic structure of the results, although not the specific optimal taxes, are unchanged by complications which occur wholly within the consumer sector. Thus, if we introduce consumption externalities that leave aggregate demand continuous we will still obtain production efficiency at the optimum, if we can argue that $V(q)$ has no unconstrained local maximum for finite q.

[10] A similar discontinuity problem arises in the case of tourists' decisions not to visit the country.

The conditions used above are no longer sufficient for this argument since the direct effects of a price change might be offset by the change in the pattern of externalities induced by the price change. Although we have not examined this case in detail, there are a number of cases where arguments similar to those in the no-externality case will be valid.[11] Furthermore it seems quite likely to us that efficiency will be desired in realistic settings.

Capital Market Imperfections

While some capital market imperfections affecting firms are complicated to deal with, some imperfections relevant only for consumers can be described as elements solely within the consumer sector. For example, consider the constraint that consumers can lend but not borrow. We must then rewrite consumer utility maximization as subject to a set of budget constraints for the different time periods. In the case of two periods, for example, it would appear as

$$(31)\qquad \begin{aligned} &\text{Maximize } u(x^1, x^2)\\ &\text{subject to } q^1x^1 + s \leqq 0\\ &\qquad\qquad q^2x^2 - s \leqq 0\\ &\qquad\qquad s \geqq 0 \end{aligned}$$

where s represents first period savings. From this consumer problem, we still have utility and demand expressible in terms of prices. We expect that the efficiency result continues to hold. In calculating the optimal formula, though, it becomes necessary to distinguish the time period of the good in question for there are now two Lagrange multipliers giving the marginal utility of income in each of the two periods. For this consumer we have

$$(32)\qquad \frac{\partial v}{\partial q_k^1} = -\alpha^1 x_k^1, \quad \frac{\partial v}{\partial q_k^2} = -\alpha^2 x_k^2$$

Since savings are allowed $\alpha^1 \geqq \alpha^2$. If the consumer would borrow if he could, $\alpha^1 > \alpha^2$ and the optimal tax structure is altered by this market limitation.

[11] We have benefited from discussions with Elisha Pazner on this subject.

REFERENCES

W. Baumol and D. Bradford, "Optimal Departures From Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *69*, 265–83.

A. Bergson, "Market Socialism Revisited," *J. Polit. Econ.*, Oct. 1967, *75*, 431–49.

M. Boiteux, "Sur la gestion des monopoles public astreints à l'équilibre budgétaire," *Econometrica*, Jan. 1956, *24*, 22–40.

G. Debreu, "A Classical Tax-Subsidy Problem," *Econometrica*, Jan. 1954, *22*, 14–22.

———, *Theory of Value*, New York 1959.

J. Drèze, "Postwar Contributions of French Economists," *Amer. Econ. Rev. Supp.*, June 1964, *54*, 1–64.

A. Prest and R. Turvey, "Cost-Benefit Analysis: A Survey," *Econ. J.*, Dec. 1965, *75*, 683–735.

F. Ramsey, "A Contribution to the Theory of Taxation," *Econ. J.*, Mar. 1927, *37*, 47–61.

P. Samuelson, "Memorandum for U.S. Treasury, 1951," unpublished.

United States Imports and Internal Pressure of Demand: 1948-68

By R. G. Gregory*

Over the past decade the successive balance of payments crises in the United Kingdom have led to considerable discussions as to the determinants of import behavior. A recurring theme in these discussions is that the demand for imports is related to the pressure on domestic resources. It is often alleged that when the rate of utilization of domestic resources is high, imports increase quite markedly (see, for example, Frank Brechling and J. N. Wolfe, and W. Godley and J. Shepherd). Although there has been very little discussion as to the theoretical link between the level of imports and the internal pressure of demand upon domestic resources, a close reading of the discussion indicates that the contributors appear to believe that the normal income and price responses of traditional theory are not sufficient to explain the cyclical behavior of imports. There is a separate response which arises from excess demand.

The recent gold crises have led to an increased interest in the *U.S.* balance of payments and a similar notion of an excess demand effect has emerged here. When excess demand occurs and capacity constraints become operative, it has been argued by W. H. Branson, and W. Lederer, E. Parrish, and S. Pizer, that imports increase quite substantially.

In this paper we look at the *U.S.* imports from the viewpoint of excess demand. We derive a neoclassical demand function for imports in Section I and then in Section II we argue that prices are slow to adjust to their equilibrium levels. Markets are cleared by a number of other variables acting as rationing mechanisms. These variables, which are typically ignored in neoclassical demand analysis, include such factors as the waiting time between the placement and delivery of an order, the alacrity with which suppliers offer trade credit, and even the general enthusiasm of the supplier to seek and obtain new orders. We argue that when there is excess demand in the United States and pressure is exerted upon domestic resources, domestic waiting times increase, credit becomes more difficult to obtain, suppliers are less vigorous in the pursuit of new orders, and consumers therefore turn to foreign suppliers. The result is an increase in imports which would not be predicted from the movement of either relative prices or domestic income. To predict the increase in imports, we need to take account of the extent to which domestic prices deviate from equilibrium as well as the relationship between income and potential output.

In Section III we confront this thesis with *U.S.* quarterly data from 1948–68. We find that there is a substantial effect on

* Senior research fellow at the Institute of Advanced Studies, Australian National University, Canberra. The paper was completed when I was visiting assistant professor at Northwestern University. My thanks to F. Brechling and colleagues of the economics department at Northwestern for encouragement and enthusiasm and to the editor for considerable advice on certain sections of the paper. The project was supported from departmental funds kindly offered to a visitor. Mr Ashley Lyman and Miss F. Wilson provided excellent research assistance. At Canberra, P. J. Lloyd has been very helpful.

imports which can be traced directly to measures of capacity constraints and excess demand. When these variables are included in the demand equation they are very significant, and the explanatory power of the demand equation is considerably increased. In the process the estimates of the conventional price elasticities are increased in magnitude and significance. Over the trade cycle the excess demand effect appears to be capable of generating a variation of 20 percent in the ratio of imports to domestic goods.

I. The Demand Equation

Most import demand equations, such as those of J. Ball and K. Marwah, and R. Rhomberg and L. Boissonneault, are not derived from utility functions and do not have microfoundations. To derive the demand equation from a utility function, as we do in this section, has a certain intuitive appeal and it does provide the step from theoretical concepts to parameter estimates. However, this approach does tend to focus attention upon a number of problems which are difficult to solve: problems of aggregation and questions such as "Whose utility function is it?" We prefer though to face these problems directly rather than specify the demand equation a priori and pretend that the problems do not exist.

The import demand equation that is derived is specified so that the dependent variable is the ratio of the *U.S.* purchase of foreign goods (imports) to the domestic purchase of goods produced in the United States. This form of demand equation assumes identical income elasticities of demand for both goods (see fn. 2) but this possible disadvantage is more than offset by the advantage to be gained when the equation is fitted to the data. By specifying the demand equation in ratio form the problem of multicollinearity, which usually plagues macro time-series analysis, is reduced and we are able to conduct much sharper tests of the numerous hypotheses that are our prime interests in this paper. Furthermore, the number of price terms which enter the demand equation is reduced to two, the price indices of home goods and imports.

We envisage the economy making decisions as to the allocation of expenditure as though it were an entity which possessed a utility function belonging to that class of utility functions which are often called utility trees. The essence of this class of utility functions is that decisions can be thought of as occurring in sequence. For our purposes we hypothesize that decisions occur at two levels. The utility function, therefore, is a two-level utility tree. The first decision is to decide upon the allocation of resources between services and goods. Once this decision has been made (the first level decision), the particular bundle of goods and the particular bundle of services must be chosen (the second level decision).

The first level utility function is assumed to be weakly separable[1] and is written as

$$(1) \qquad U = F[\phi_1(X^{(1)}), \phi_2(X^{(2)})]$$

where $X^{(1)}$ and $X^{(2)}$ are vectors of goods and services, respectively. In this paper we are primarily concerned with the second level of the first branch of the utility function; that is, our interest lies in the allocation of resources between goods, and in particular, the allocation between goods produced at home, $X_1^{(1)}$, and goods imported from overseas, $X_2^{(1)}$. The second level utility function which relates these two goods we write as

[1] A utility function is weakly separable with respect to the partition $\phi_1(X^{(1)})$, $\phi_2(X^{(2)})$ if the marginal rate of substitution $u_i(X)/u_j(X)$ between two commodities i and j from $\phi_s(X^{(s)})$ is independent of the quantity of commodities outside of $\phi_s(X^{(s)})$. (See S. M. Goldman and H. Uzawa.)

$$\phi_1(X^{(1)}) = \left[\sum_{i=1}^{2} \delta_i (X_i^{(1)})^{-p} \right]^{-1/p},$$

(2)

$$\delta_i > 0, \quad -1 < p = \frac{1-\sigma}{\sigma} < \infty$$

which is the familiar constant elasticity of substitution form.

The problem then is to maximize equation (2) subject to the budget constraint

$$Y^{(1)} = \sum_{i=1}^{2} p_i^{(1)} X_i^{(1)} \tag{3}$$

where $p_i^{(1)}$ are prices of $X_i^{(1)}$. $Y^{(1)}$, which is predetermined from the first level allocation problem, is the quantity of resources to be spent on the goods $X_i^{(1)}$. The necessary condition for a solution to this problem is given by the equality of the marginal utility ratio with the ratio of prices, that is,

$$\frac{X_2}{X_1} = \left(\frac{\delta_2}{\delta_1}\right)^{\sigma} \left(\frac{p_1}{p_2}\right)^{\sigma} \tag{4}$$

As we are only concerned with the first branch of the utility function (1), we write the variables of equation (4) and the following expressions without superscripts unless there is some doubt as to their identity.[2]

As a further refinement we may wish to allow for a change in tastes over the data period. This can be done without altering the basic utility function, or the magnitude of its parameters, if it is assumed that tastes change so as to be X_1 or X_2 augmenting. If the change in tastes is X_1 or X_2 augmenting, it can be entered into the utility function as

$$\phi_1(X^{(1)}) = \left[\sum_{i=1}^{2} \delta_i^{(1)} (\eta_i^{(1)} X_i^{(1)})^{-p} \right]^{-1/p} \tag{5}$$

where η_i, are the change in taste multipliers. This specification of the effect of a change in tastes will enable us to measure the relative change in tastes between foreign and domestic goods but not the absolute change in satisfaction derived from each individually.[3] The notion of a

---

[2] Some of the more obvious properties of this theoretical framework should be mentioned. The advantage, for empirical application, to be gained from writing (1) as weakly separable and (2) as homogeneous is that the second level allocation decision, X_2/X_1, is independent of the first-level decisions and, as is evidenced by (4), this simplifies the estimation problem considerably. Only two price indices enter (4) and consequently the allocation of resources between the two goods, X_1 and X_2, is independent of the price of services. The Slutsky income-compensated elasticity of demand for any good, however, does depend upon the first level decisions. This price elasticity, $E_{ii}^{(1)}$ can be written as,

$$E_{ii}^{(1)} = \frac{v_j^{(1)}}{v_i^{(1)} + v_j^{(1)}} \sigma_{ij}^{(1)} - \frac{v_i^{(1)}}{v_i^{(1)} + v_j^{(1)}} E_{is}^{(2)}$$

where v_j is the expenditure on $X_j^{(1)}$, $E_{ii}^{(1)}$ is the own price elasticity of demand for the ith good and $\sigma_{ii}^{(1)}$ is the elasticity of substitution between the two goods. Since the mean proportion of goods imported over the data period, $v_i^{(1)}/v_i^{(1)}+v_j^{(1)}$, is approximately 13 percent the second term can be overlooked if $E_{is}^{(2)}$, the elasticity of the ith good with respect to a change in the price of services, is not too large. In these circumstances the elasticity of demand for imports can be approximated by σ_{ij}.

Another important property of this theoretical framework is that, since (2) is homogeneous, the allocation of resources between the two goods, X_1 and X_2, is independent of the quantity of resources, $Y^{(1)}$, devoted to them. Consequently, this implies $I_i^{(1)} = I^{(1)}$ where I is each respective income elasticity; that is, the income elasticity of each good separately, and the income elasticity of the composite index of goods, are equal. There is no other restriction placed upon the income elasticities. For a further discussion of two-level functions see K. Sato where the concept is applied to production functions.

[3] If it is assumed that the multipliers η_i change at a constant exponential rate with time, that is,

$$\frac{\eta_1}{\eta_2} = \eta_0 e^{(r_1 - r_2)t} = \eta_0 e^{rt}$$

where r_i is the change of tastes associated with the ith good, then upon substitution into (6) it will be an easy matter to estimate the relative change of tastes. The coefficient on the time trend in the demand equation will be $(1-\sigma)\lambda(r_1-r_2)$. If the coefficient on the time trend is estimated to be positive, and $(1-\sigma)$ and λ are also positive (the adjustment specification requires $0<\lambda<1$), it follows that r_1 is greater than r_2 and, *ceteris paribus*, the change in tastes has increased the marginal utility of X_2 relative to X_1. If $(1-\sigma)$ is negative the opposite conclusion would be reached. The in-

change in tastes augmenting the quantity of a good consumed has an analogous counterpart in the theory of production where factor augmenting technological change has been used in a number of studies to measure the sources of output growth, (see, for example, P. A. David and Th. Van De Klundert).

Finally, if we allow for a delayed response arising from costs of adjustment to changes in relative prices on the part of the consumer we can add a Koyck lag[4] to the demand equation and write it as,

$$\frac{X_2}{X_1} = \left(\frac{\delta_2}{\delta_1}\right)^{\sigma\lambda}\left(\frac{\eta_1}{\eta_2}\right)^{(1-\sigma)\lambda} \cdot\left(\frac{P_1}{P_2}\right)^{\sigma\lambda}\left(\frac{X_2}{X_1}\right)_{t-1}^{(1-\lambda)} \quad (6)$$

This concludes the derivation of the basic demand equation. In the next section the demand equation is considered in more detail in order to show that it may not focus attention upon all the important variables which affect the decision to buy foreign or domestic goods.

II. The Effective Price Relationship

The demand equation derived in the previous section focuses attention upon relative prices, tastes, and a lag in the adjustment to the equilibrium quantities, as the demand variables and relationships which are important in determining the allocation of resources between foreign and domestic goods. However, there is some evidence that actual prices are slow to adjust to their equilibrium values and that in the *short run*, markets are cleared partly in response to relative waiting times, trade credit terms, rebates, discounts, and the general ability of the sellers to meet customer requirements. For example, in a recent paper, M. D. Steuer, R. J. Ball, and J. R. Eaton (Nov. 1966), have shown that *independent of relative prices*, the relative waiting time between the order and delivery of machine tools is a crucial rationing variable which plays an important part in allocating the flow of new orders among the machine tool industries of the United Kingdom, United States, and West Germany. In another paper (see Ball et al. Sept., 1966), they show that, *independently of relative prices*, capacity constraints affect the level of exports of manufactured goods from the United Kingdom. The implication of both these studies is that markets are being cleared by other variables as well as the usual price variable that is included in demand studies.

If this phenomenon is widespread then conventional economic analysis, which focuses upon income and relative prices as the explanatory variables in a demand equation, may, for particular problems, be quite misleading. In econometric work, for example, the omission of relevant variables from a demand equation can lead to biases and poor predictive performances since, at any given price, different quantities may be demanded depending upon the magnitude of the omitted variables. The same problems may arise with respect to supply curves.[5]

terpretation of a given sign of the coefficient on the time trend therefore depends upon whether σ is greater or less than unity.

[4] The adjustment model hypothesized is

$$y_t - y_{t-1} = \lambda(y_t^* - y_{t-1})$$

where y_t is *log* (X_2/X_1) and y_t^* is the *log* of the equilibrium ratio of consumption after consumers have fully adjusted to the prevailing relative prices.

[5] The notion that empirically there may be no unique short-run demand or supply curves when these are defined in the traditional neoclassical manner is gradually becoming more common. Otto Eckstein and Gary Fromm, for example, in their recent paper in which they discuss the determinants of price, refer to "disequilibrium demand and supply curves" and suggest that disequilibrium rather than equilibrium is the more common situation. If this is an accurate description of markets then most data observations will not lie on either a neoclassical demand or supply curve. If the price is fixed, and there is excess demand, then the gap between the demand and supply curve defines the area

If it is true that in the short run other variables, rather than price, predominantly clear markets, and that disequilibrium situations are more common than equilibrium situations, then there is considerable theorizing yet to be done to specify the dynamic interrelationship between each of the market clearing variables, the relationship between these variables and demand and supply curves, and to analyze in general the economics of disequilibrium situations. Furthermore, for empirical purposes the formulation of demand and supply curves should explicitly take account of disequilibrium situations and introduce each of the market clearing variables explicitly into the functions.

Definition of the Effective Price

Perhaps the simplest modification that can be made to neoclassical theory to meet the criticisms above is to redefine the price variables of equation (6) as *effective prices.* Rather than treat the price of a commodity as a one dimensional variable in the usual textbook manner, it may be more appropriate to define the price as a vector possessing many dimensions. We call this vector the effective price, P_e. Its elements are the actual quoted price P_a, the waiting time, the trade credit terms, rebates and any other ancillary aspects of the contract which are relevant for the decision to purchase or not.[6] Consideration of *relative effective prices* at home and abroad determines whether the commodity is purchased from domestic or foreign suppliers. From this point onwards we conduct the analysis in terms of the domestic effective price but it should be remembered that a similar analysis applies to foreign effective prices and that it is relative effective prices that should enter the demand equation.

If each of the ancillary aspects of contracts are defined as ω_i the effective price could be written as,

$$(7)\quad P_e = APa\prod\omega_i^{\gamma_i} \qquad \gamma_i \gtrless 0,\ i = 1 \ldots N$$

where γ_i are the exponential weights attached to each of the ω_i elements. The equation has been normalized so that the exponent on the actual price term is unity. For elements such as the waiting time the coefficient γ is positive so that an increase in the waiting time between the placement of an order and the delivery of the goods is equivalent to a price increase. Other elements of ω, such as the willingness of the seller to adjust to the requirements of the buyer will possess negative γ exponents. From the buyer's point of view an increase in the willingness of the seller to adjust to his demands is equivalent to a price reduction.

If each of the elements of the effective price are observable, independent of each other, and exogenous to demand decisions, then the effective price ratio could be substituted into (6) for the actual price ratio and the demand equation estimated. However, not all the elements of the vector ω are observable. In some instances, such as the ease at which trade credit is obtainable, the elements are measurable in principle but data series do not exist. In other instances, such as the general enthusiasm and desire on the part of sellers for new orders, the elements are not directly measurable. Consequently, we need to develop proxy variables to measure the elements of the effective price vector. This is done in the next two subsections. First,

in which the price quantity observation must fall. Its exact location will depend upon the extent to which suppliers deviate from the equilibrium supply curve by drawing down inventories and temporarily increasing production. A brief survey of some of the recent literature which expounds similar views in the fields of inflation and employment theory can be found in Edmund S. Phelps.

[6] An alternative way of thinking about this problem is to follow the approach adopted by K. Lancaster and consider the ancillary aspects of contracts as characteristics of the goods.

we discuss in some detail the derivation of a variable to measure the quoted waiting time between the placement and delivery of an order. This variable will probably be an important component of the effective price. Then a more general theory is developed to relate the elements of ω to observable economic variables. This theory should be regarded as suggestive of possible relationships to enable the development of suitable proxy variables rather than a formally complete model.

Measurement of the Quoted Waiting Time

The waiting time which is relevant for the demand equation is not the actual *ex post* waiting time but the *ex ante* waiting time which is quoted to the prospective customer. The *ex post* waiting time is likely to differ from that quoted if it is uneconomic to calculate the waiting time for each individual customer and if the production flow is subject to some inflexibility. Since these circumstances are likely to exist in most industries the suppliers may form their quotations lq, by using the *ex post* waiting time of the previous period, la, adjusted for its rate of change. In that case, the quoted waiting time in any period might be expressed as,

$$lq = la_{-1}\left(\frac{la_{-1}}{la_{-2}}\right)^{x} \qquad x > 0 \tag{8}$$

where x is a positive exponential weight attached to the rate of change of the waiting time and the subscripts refer to the time period. If the waiting time has been increasing, $la_{-1} > la_{-2}$, then suppliers will take account of this fact and quote a waiting time which exceeds that of the previous period. Similarly, if the waiting time has been decreasing the quoted waiting time will be less than that of the previous period.

There are a number of interesting implications both of theory and of interpretation of the regression equations which might be developed from this specification of the waiting time variable. Ignoring the foreign waiting time for the moment, the domestic quoted waiting time can be included into the demand equation by substituting (8) into (7) and then substituting (7) into (6) to derive,

$$\begin{aligned} ln\left(\frac{X_2}{X_1}\right)_t = B' + \lambda\sigma ln\left(\frac{Pa_1}{Pa_2}\right)_t & \\ +\gamma\lambda\sigma(1+x) ln\, la_{t-1} & \\ -\gamma x\lambda\sigma ln\, la_{t-2} & \end{aligned} \tag{6'}$$

where,

$$B' = ln\left[A^{\lambda\sigma}\left(\frac{\delta_2}{\delta_1}\right)^{\lambda\sigma}\left(\frac{\eta_1}{\eta_2}\right)^{(1-\sigma)\lambda}\left(\frac{X_2}{X_1}\right)^{1-\lambda}_{t-1}\right]$$

and γ is the positive exponential weight that is attached to the quoted waiting time element of the effective price. The empirical results from applying (6′) to the data should give a positive coefficient, $\gamma\lambda\sigma \cdot (1+x)$, on the first actual waiting time variable, a negative coefficient, $-\gamma x\lambda\sigma$, on the second waiting time variable and an absolutely larger coefficient on the first waiting time variable than on the second, $|\gamma\lambda\sigma(1+x)| > |-\gamma x\lambda\sigma|$. Furthermore, by taking the estimate of $\lambda\sigma$ from the coefficient on the relative actual price term the parameters γ and x, which should both be positive, can be identified.

Data on the *ex post* waiting time for goods produced in the United States are not available but the waiting time can be approximated[7] by dividing the level of unfilled orders, uo, by the current production rate, Q, so that

$$la \simeq \frac{uo}{Q} \tag{9}$$

[7] If there is little variance in the rates of growth of production and new orders then this approximation is reasonably accurate. The degree of accuracy worsens if either series fluctuates.

The substitution of (9) into (8) completes the derivation of lq which is an estimate of one of the elements of the vector ω.

Since the above technique to estimate the quoted waiting time may be inaccurate and/or the relationship between the various elements of the ω vector may be too complex for their variation to be approximated by the quoted waiting time, we now attempt to derive a number of proxy variables from a more general theory of the relationship between the elements of the effective price vector and observable economic variables.

A Theory of the Effective Price Relationship

First we assume that firms are imperfect competitors in the product market in the sense that they can control P_a as well as all the elements of ω. They supply only the domestic market.

Second, to develop the theory as simply as possible we assume that the capital stock of each firm is fixed. We do, however, allow for variations in the inventory level of the firm.

Third, we define ω^* to be the long-run equilibrium value of the vector ω. In this paper our primary interest is in disequilibrium situations so ω^* will be regarded as a constant.[8] Likewise I^* is the long-run equilibrium level of inventories which, at this stage, is also regarded as a constant.

Fourth, there is a cost function defined upon different levels of ω such that the further ω departs from ω^* the greater the cost the firm must bear. If ω exceeds ω^* and waiting times are too long and credit terms too stringent then customer goodwill and future orders will be lost. If ω is less than ω^* and waiting times are too short and credit terms too cheap then production planning may be more difficult and the amount of capital required to finance the trade credit more than optimal. The cost of deviations of ω from ω^*, however, is always less than that of not clearing the market. We would argue that the loss of goodwill that occurs when a firm simply refuses to supply the product at the current price and current waiting time exceeds the loss of goodwill which may occur if the waiting time or other elements of ω are increased. Similarly, there is a cost function defined upon different levels of I such that the further I departs from I^*, the greater the cost the firm must bear.

From these assumptions the demand, Q_D, for the product of the typical firm can be written as,

$$Q_D = g(P_a, \omega) \tag{10}$$

and its supply curve[9] defined as

$$Q_S = h(\bar{K}, T, W, g(P_a, \omega), \omega, I) \tag{11}$$

where $\bar{K}$ is the fixed quantity of capital, T the level of technology, W the wage rate of labor, and I the inventory level.

Fifth, we assume that firms are reluctant to change P_a frequently in the short run. Their reluctance may be rationalized by postulating that the administrative costs of changing the price are considerable and that frequent price changes destroy the goodwill of customers. In large companies with many products, many outlets, and a large sales force it may even be physically impossible to change prices in immediate response to changes in demand and supply conditions.

Sixth, since firms are reluctant to change P_a in the short run, markets will be cleared

[8] In a long-run context, ω^* might be an endogeneous variable. There is some evidence that some of the elements of ω^* differ across countries. Steuer, Ball, and Eaton (Nov., 1966) discovered that the United Kingdom is known as a "low price-long delivery lag" supplier of machine tools in contrast to the United States which is a "high price-short delivery lag" supplier.

[9] The supply equation (11) states that the quantity supplied which will maximize profits can be calculated from the demand curve and those factors which determine marginal cost.

by other devices so that any demand at current effective prices will be satisfied. Consider, for example, a situation in which equilibrium is disturbed by a once and for all injection of excess demand, Z. Excess demand either positive or negative, is defined to exist when

$$(12)\ Z = Q_D^* - Q_S^* = g(Pa, \omega^*) - h(\bar{K}, T, W, g(Pa, \omega^*), \omega^*, I^*)$$

is either positive or negative. This definition of excess demand is not quite parallel to the usual definition because here excess demand is measured as the gap between the quantity demanded and supplied at current prices but *equilibrium* values of ω and I. When equilibrium is disturbed by a once and for all injection of excess demand, it is hypothesized that the responses of the firm can be divided into three separate groups and that these groups of responses follow each other sequentially through time. Within each group, however, it is assumed that the responses occur simultaneously. These sets of decisions consist first of an inventory response, than an ω, output, and a further inventory response, and finally a price response. We now turn to a further elaboration of this behavior pattern. Not all the disequilibrium relationships will be developed, only those which enable ω to be measured by variables for which there are existing data series.

The first response to a once and for all injection of excess demand, as defined above, is that the inventory level changes to meet the gap between the production rate and the quantity demanded. This response can be written as,

$$(13)\ \frac{I_{-1}}{I^*} = j(Z_{-1});\ \text{where}\ \frac{I_{-1}}{I^*} \gtrless 1 \text{ as } Z_{-1} \lessgtr 0$$

I^* is the equilibrium level of inventories which is assumed to be a constant, and the injection of excess demand initially occurs in period t−1. There are a number of reasons why this might be the first response. In most firms there will be both a lag between the advent of excess demand and its recognition and lag between that recognition and action. During this time the inventory level will alter as a result of the excess demand conditions. Furthermore, in most firms the costs of adjustment attached to the change in the inventory level may be small compared to the costs of adjustment attached to the other economic variables. If the excess demand persists the inventory level continues to fall and the costs imposed upon the firm increase. It then becomes optimal to adjust some of the other variables.

The second set of responses consist of three actions which are mutually determined. These involve a further change in the inventory level, an output response, and a change in the effective price by altering all or some of the elements of ω. The functional relationship which defines the combinations of these responses which will meet the excess demand can be written as,

$$(14)\qquad k\left(\frac{I}{I_{-1}}, \frac{\omega}{\omega^*}, \frac{Q}{Q_{-1}}; Z_{-1}\right) = 0$$

where Q is the rate of output. Equation (14) describes a behavioral relationship between these three endogenous variables and excess demand. For any given Z_{-1}, equation (14) can be thought of as a three dimensional surface, concave to the origin, where for ease of exposition the axes are defined as the ω response, the output response, and the *inverse* of the inventory response. This surface connects together all those points which would ensure that demand is satisfied. For movements around this surface the partial derivatives of each of the ω, output and *inverse* of the inventory responses with respect to each other and keeping Z_{-1} fixed are all negative. For higher levels of Z_{-1} the surface moves outwards from the origin.

The combination of these three re-

sponses is chosen by minimizing, subject to (14), the sum of the cost function defined upon the level of output and the cost functions defined upon the deviations of ω and I from their equilibrium levels. It is assumed that (14) and the sum of the cost functions are such that the expansion path, for different levels of Z_{-1}, is monotonic; that is, larger increases in output and ω are always associated with greater reductions in inventories. This expansion path can be written as,

$$\omega = m\left(\frac{I}{I_{-1}}, \frac{Q}{Q_{-1}}\right) \tag{15}$$

where the association between changes in ω and changes in inventories is negative and between changes in ω and changes in output positive. Since we are concerned with short-run responses, ω^* and I^* are regarded as constants and omitted from the functional specifications. This completes the description of the second set of responses.

Since the equilibrium values of I and ω are assumed to be constant the level of excess demand, as defined above, and hence long-run disequilibrium, will persist despite this second set of responses. This situation leads to the third set of responses. The third response is to change the price level, thereby remove excess demand, and to bring the ω elements, inventories and the output rate to long-run equilibrium levels. For our purposes, only the price response is needed and it can be written as,

$$\frac{P_{+1}}{P} = n(Z_{-1}), \tag{16}$$

where

$$\frac{P_{+1}}{P} \gtrless 1 \text{ as } Z_{-1} \gtrless 0$$

This expression, which has a long history in discussions of excess demand (see, for example, Paul Samuelson), completes the model.

We are now in a position to derive proxy variables for the vector ω which can be substituted into the effective price relationship (7) which in turn can be substituted into the demand equation (6). Equation (15) provides one set of proxy variables and if the theory outlined above were correct then this is all that is needed. However, given the degree of complexity of the dynamic interrelationships of the various responses perhaps the best strategy, at this stage, is to derive other sets of proxy variables to provide further tests of the theoretical framework. This can be done in two different ways. The first and simplest is as follows. From (15) it is known that ω and excess demand Z_{-1} are positively associated. From (13) and (16) it is known that inventory reductions in period t−1 and price increases in period t+1 are also positively associated with excess demand. Therefore, ω could be measured in terms of these variables,

$$\omega = j'(I_{-1}) \tag{13'}$$

or

$$\omega = n'\left(\frac{P_{+1}}{P}\right) \tag{16'}$$

The monotonicity of the original relationship will give unambiguous sign associations between these proxy variables and ω. There is a positive association between the ω elements of the effective price and the future rate of change of prices and a negative association between the ω elements of the effective price and the past change of inventories from their normal level.

The second and more complex procedure is to use (14) to substitute for any of the variables on the right-hand side of (15) (thus introducing Z_{-1} explicitly into (15)) and then use (13) or (16) to substitute for Z_{-1} to introduce into the range of proxy

variables price and inventory changes as well as one of the original variables on the right-hand side of (15). It is fairly easily shown, however, that this technique does introduce some ambiguity in the sign of the coefficient of the original variable which appeared in (15).

As an example of this procedure we can use (14) to derive Q/Q_{-1} in terms of I/I_{-1}, ω and Z_{-1} and then substitute this relationship for Q/Q_{-1} into (15). Then for Z_{-1} we can substitute either the inverse of (13) or (16), and rearrange the expression to give, respectively:

$$(15') \qquad \omega = m'\left(\overset{?}{\frac{I}{I_{-1}}}, \overset{-}{I}_{-1}\right)$$

or

$$(15'') \qquad \omega = m''\left(\overset{?}{\frac{I}{I_{-1}}}, \overset{+}{\frac{P}{P}}\right)$$

where the symbols above the variables define the predicted sign pattern of their coefficients.[10] Although the sign of particular coefficients may be ambiguous the theory often predicts relationships between them. For example, if (15′) is specified as a log-linear function it becomes

$$(17) \qquad \log \omega = \alpha_1 \log \frac{I}{I_{-1}} + \alpha_2 \log I_{-1}$$

or rearranging

$$(17') \qquad \log \omega = \alpha_1 \log I + (\alpha_2 - \alpha_1) \log I_{-1}$$

From (15′) above it is known that α_1 is of ambiguous sign but that α_2 is negative. Consequently if α_1 is positive then $|\alpha_1| < |\alpha_2 - \alpha_1|$. If α_1 is negative then $(\alpha_2 - \alpha_1)$ cannot be greater than $-\alpha_1$ without violating the restriction that α_2 be negative. Likewise, if (15″) is specified as a log-linear relationship and the rate of change of inventories entered in the regressions as the inventory level and its lagged value, then the signs of the coefficients should alternate although there is no restriction on whether the first coefficient is positive or negative.

To conclude this section, the relationship between the effective price and excess demand on the one hand, and the interrelationship between the elements of the effective price on the other, are likely to be more complex than the simple theoretical relationships sketched above. It is conceivable that each measure of the ω vector that we have discussed will capture a different aspect of these relationships. In the next section therefore we will experiment with different combinations of these measures.

III. The Empirical Results

In the first part of this section the model is used to explain the quarterly variation of the imports of general merchandise into the United States from 1948 to 1968. Total imports, of course, contain a significant proportion of raw materials and producer goods,[11] both of which may not at first glance seem to fit easily into the demand framework developed in Section I. However, I feel justified in choosing this level

[10] As these substitutions are a little complicated it may be useful to spell the procedure out in a little more detail. Consider the derivation of (15′). We can rearrange (14) to write

$$(1') \qquad \frac{Q}{Q_{-1}} = r\left(Z_{-1}, \omega, \frac{I}{I_{-1}}\right) \qquad r_1, r_3 > 0, r_2 < 0$$

Substituting (1′) into (15) gives

$$(2') \qquad \omega = s\left(\overset{-}{\frac{I}{I_{-1}}}, \overset{+}{Z}_{-1}, \overset{-}{\omega}, \overset{+}{\frac{I}{I_{-1}}}\right)$$

where the associations between the variables on the right-hand side and ω are given above the variable. The important fact to notice is that the first and fourth term of (2′), I/I_{-1}, possess different signs. Therefore, inverting (13), substituting for Z_{-1} in (2′) and then rearranging we derive (15′) in the text.

[11] If we were primarily interested in raw materials or producer goods the utility functions of Section I would be replaced on *CES* production functions. There is no a priori reason why production functions should not consist of more than one level. The article by Sato contains a full discussion of two-level production functions.

of aggregation because it is of direct policy interest and, more importantly, the general notion of an effective price is applicable to *all* goods that are obtainable from both foreign and domestic suppliers. In a later part of this section, I briefly examine more disaggregated data. I also consider a simpler model which measures capacity constraints directly.

The Imports of General Merchandise

This section presents the results of a number of regressions which include different proxy variables to measure the effective price. It is found that, independently of the choice of the particular set of proxy variables suggested in Section II, the concept of a domestic effective price proves to have considerable ability to explain the behavior of the ratio of *U.S.* imports to the domestic production of goods. All coefficients possess the correct sign and almost all are significant at conventional levels. Insignificant coefficients are occasionally encountered when the rate of production is used as one of the proxy variables to measure the ω elements.[12] These results are not reported. All other coefficients of proxy variables are very significant.

The only failure of the model is associated with the measurement of the foreign effective price. As a result of data difficulties the proxy variables used were very different from those which were used to measure the domestic effective price. At this time I have had no success with any of these variables and they are excluded from the published results.[13]

Table 1 lists a selection of the results. In all regressions the R^2 are between .92 and .95. The first equation of Table 1 is the standard import function (6) where the price term is expressed as the relative actual[14] price rather than the effective price ratio. This equation does not perform as well as those that include the effective price as the independent variable. Furthermore when the effective price is included as an independent variable there are a number of common features running through each of the regressions.

First, the short-run elasticity of substitution, with respect to changes in the actual price ratio, increases quite considerably although the new estimate depends upon which specification of the effective price is used. The short-run elasticity of substitution, which is approximately equal to the own price elasticity of demand for imports (see fn. 2), is usually less than unity but the long-run price elasticity always exceeds unity. The typical value of the long-run elasticity is in the vicinity of three which is a much larger estimate than has been found in earlier studies (see, for example, Ball and Marwah).

Second, the significance level of the actual price elasticity increases quite

[12] The referee has pointed out that among the third set of responses, which have not been developed in Section II, there may be a further rise in output accompanying the fall in ω. Consequently unless the sets of responses developed in Section II coincide with our data time periods the variable Q/Q_{-1} may well be a bad proxy variable for ω.

[13] At this juncture this should not be too serious a source of concern. When all foreign countries which supply the United States with imports are grouped together then, at any point of time, some will be experiencing positive excess demand and others negative excess demand so that their aggregation will involve some cancelling. Furthermore, the proxy variables used are, on a priori grounds, not as satisfactory as their *U.S.* counterparts. I have tried the rate of change of import prices, the deviation of world imports around their trend and the deviation of European production around its trend.

[14] As a price index for imports I used either the implicit price deflator of imports or a unit value index. Both were taken from the *Survey of Current Business* and are subject to a number of special difficulties, see, for example, the Joint Economic Committee *Hearings* 1961. The implicit price deflator usually gave slightly higher estimates of the elasticity of substitution. The price index of imports used in Table 1 is the implicit price deflator of imports.

TABLE 1—THE RATIO OF IMPORTS TO DOMESTIC GOODS PRODUCTION REGRESSED AGAINST VARIOUS SPECIFICATIONS OF THE EFFECTIVE PRICE RATIO
Quarterly Data 1948–68 (*t*-values in parentheses)

Equations	1	2	3	4	5	6
Constant	.4800 (2.85)	.8528 (5.13)	1.0683 (6.31)	.8731 (2.90)	1.2316 (4.21)	.7149 (4.19)
$\frac{P_1}{P_2}$	.2308 (2.10)	.8906 (4.68)	1.1551 (5.85)	.8478 (4.55)	1.1412 (6.02)	.4311 (3.68)
t	.0009 (1.86)	.0012 (2.63)	.0011 (2.56)	.0010 (2.30)	.0010 (2.32)	.0010 (2.21)
$\frac{X_{2t-1}}{X_{1t-1}}$	.8075 (11.54)	.6889 (10.44)	.6097 (9.20)	.7180 (10.40)	.6232 (9.15)	.7143 (10.15)
l_{at}		.5712 (5.00)	.4764 (4.26)	.5213 (4.22)	.3989 (3.37)	
l_{at-1}		−.4340 (4.51)	−.3246 (3.38)	−.3747 (3.28)	−.2288 (2.06)	
$\frac{P_{1t}}{P_{1t-1}}$			1.3971 (2.26)		1.3353 (2.25)	1.9064 (3.01)
$\frac{P_{1t-1}}{P_{1t-2}}$			1.7333 (2.49)		2.0104 (2.97)	1.2780 (1.82)
$\left(\frac{I}{X_1}\right)_{t-1}$				−.6479 (2.32)	−.7559 (2.94)	
$\left(\frac{I}{X_1}\right)_{t-2}$				.5789 (2.28)	.6084 (2.62)	
S.E.R.	.048	.042	.039	.040	.038	.044
R^2	.92	.94	.95	.94	.95	.93
D	1.34	1.25	.68	1.87	1.21	.23

Symbols: X_1 = the domestic production of goods, constant prices; X_2 = the value of general imports, constant prices; P_1 = implicit price deflator of domestic goods output; P_2 = implicit price deflator of imports; l_a = average delivery lag; t = time; I = manufacturing and trade inventories.

Source: Data taken from the *Survey of Current Business; D* is a new statistic devised by J. Durbin to test for serial correlation when the equation contains lagged dependent variables. All variables are expressed as logarithms, seasonally adjusted, and measured in constant prices.

considerably[15] as proxy variables for ω are added to equation (6). In almost all instances the level of significance is doubled.

Third, the coefficient on the time trend is not very sensitive to the way in which the effective price is measured. Since the coefficient on the time trend is positive,

[15] Both these results are to be expected. If the demand equation is misspecified to include only the actual price as a measure of the effective price variable then this misspecification is similar to an errors in variable problem. The omitted deviations of ω around ω^* are very similar to a random measurement error. Under these circumstances, as elements of the effective price are added to the regression equation the price variable is gradually purged of its errors and the downward bias is removed. Furthermore, the significance level should increase as the actual price variable is allowed to capture more and more of its true effect.

and the long-run elasticity of substitution is always greater than unity, the time trend coefficient implies that tastes have been changing towards foreign goods (see fn. 3). The effect of the implied bias of tastes towards foreign goods, however, is always small, being somewhere between $-.03$ and $-.002$ percent per quarter.[16]

From the preceding paragraphs it is clear that the substitution of the effective price for the actual price has led to a significant improvement in the demand equation. We now turn to an analysis of the various specifications of the effective price.

Consider the waiting time variables first. In Section II, it was suggested that the quoted waiting time would depend upon the actual waiting time and its rate of change. The implications of this hypothesis (the reader should refer back to (6′)) are, a positive coefficient on the waiting time variable, a negative coefficient on the waiting time variable lagged one quarter, and that the absolute value of the positive coefficient should exceed that of the negative coefficient. The results of each equation in Table 1 satisfy these restrictions. As further variables are added to the regression equation 2, the only discernible effects upon the waiting time coefficients are a slight reduction in their absolute magnitude and significance. The coefficients always remain significant at conventional levels. In the regression equations we found that the current waiting time and the waiting time lagged one quarter (see (6′)) performed better than the waiting time lagged one and two quarters, respectively. This should only give rise to some concern if there is some a priori reason why the theoretical time periods of Section II should exactly coincide with quarters.

As a further step we can decompose the estimated coefficients on the waiting time variables to derive estimates of the exponent, γ, and the weight, x, which is attached to the rate of change effect[17] (see equations (7) and (8)). We find that γ is approximately .15. This implies, for example, that a 50 percent increase in the quoted waiting time is equivalent to an 8 percent increase in the actual price.[18] The estimates of x lie between 2.5 and 3.0. These estimates imply, for example, that a 20 percent change in the actual waiting time from one quarter to the next gives rise to an 80 percent increase in the quoted waiting time.

The third regression equation includes the quoted waiting time, (see equation (8)) and as another measure of the ω elements of the effective price, the rate of change of the implicit home goods deflator, equation (16′). The theory of Section II predicts a positive coefficient on the rate of change of price variable. In the regressions reported, the rate of change of prices for two quarters has been included. That both are positive and significant might be explained by the aggregation of different industries

[16] The coefficient on the time trend is $\lambda(1-\sigma)r$. If we use equation (1) as an example we can solve for r as follows: $\lambda \cong .2$, $\sigma \cong 1.15$ and $\therefore -.2(.15)r = .0009$ which gives r equal to $-.03$ percent per quarter which means that $|r_2|$ exceeds $|r_1|$ and tastes have changed towards foreign goods. If, however, there is a time trend in the inventory-output ratio then the coefficient in the time trend would include this effect and r can no longer be identified.

[17] As an example of these derivations consider regression equation (4) where $\lambda\sigma \cong .84$, $\gamma\lambda\sigma(1+x) \cong .52$, $-\gamma\lambda\sigma x \cong -.37$. Consequently, $\gamma\lambda\sigma \cong .52 - .37$ and $\gamma \cong \frac{.15}{.84} \cong .17$. To derive x we have $\gamma\lambda\sigma x \cong .37$, and therefore $x \cong \frac{.37}{.15} \cong 2.6$.

[18] Taking the differential of (7) with respect to P_a and l_q gives,

$$dPe = \gamma Pe \frac{dl_q}{l_q} + Pe \frac{dP_a}{P_a}$$

Setting dPe equal to zero, rearranging and substituting the estimate of γ from equation (2), (see fn. 17) gives

$$\frac{dP_a}{P_a} = -.17 \frac{dl_q}{l_q}$$

and therefore a 50 percent increase in the quoted waiting time is equivalent to an 8.5 percent increase in the actual price.

in which the lags between the relationships developed in Section II differ. In conclusion, the greater the rate of change of prices over the last two quarters, the greater the value of the ω elements during and before those quarters; and consequently, given the lag between the placement of an import order and its arrival, the greater the value of current imports.

The fourth equation includes the level of inventories and its rate of change. Inventories have been normalized by dividing through by the production rate. The coefficients possess the correct sign pattern (see the discussion of equation (17′) in Section II) and are both significant. The condition derived in Section II that if α_1, the first coefficient, is negative then $(\alpha_2-\alpha_1)$, the second coefficient, should not be greater than $-\alpha_1$, is also satisfied.

The fifth equation includes the quoted waiting time and the variables of equation (15″) as a measure of the ω elements. All coefficients again satisfy the restrictions of Section II and are at the same time significant at conventional levels. The coefficients on the inventory variables should be approximately of equal magnitude but opposite in sign and the coefficients on the rate of change of prices positive.

The sixth equation omits the quoted waiting time and measures the ω elements by an approximation to equation (16′). Again the coefficients possess the correct sign and are significant. We now turn to a discussion of the adjustment speeds implied by these equations.

Each of the equations in Table 1 contains a lagged dependent variable. The advantage to be gained from this specification is that it enables an estimate to be made of the lagged response of consumers to changes in the effective price ratio. The estimates of the coefficients on the lagged dependent variable imply an average lag of one and a half to four quarters and a long-run price elasticity, which is two to three times larger than the short-run price elasticity. We also find that as the specification of the demand equation improves, that is as proxy variables are included to measure the effective price, then the estimate of the response speed is increased. The average lag is reduced from four quarters (equation (1)) to approximately one and a half (equations (3) and (5)). The use of lagged dependent variables, however, does introduce a number of problems which may give rise to some concern.

It has been shown by E. Malinvaud and others that when there is both serial correlation amongst the error terms and a lagged dependent variable then the application of least squares regression will bias the coefficient on the lagged dependent variable either up or down depending on whether the serial correlation is negative or positive. The estimate of the adjustment speed and the long- and short-run multipliers will therefore be biased. Furthermore, the estimate of the variance of the coefficient on the lagged dependent variable will also be biased but the evidence that is available suggests that this bias will be small (see, for example, Malinvaud). It is therefore of some interest to know whether serial correlation is present in the equations of Table 1.[19]

[19] A somewhat related problem to this is that the inclusion of lagged dependent variables may well give rise to misspecification problems. Z. Griliches has pointed out that if the correct specification of a model includes serial correlation of the error term but excludes the lagged dependent variable then the inclusion of a lagged dependent variable and the application of least squares will usually result in an improved fit to the equation and a significant estimate of the coefficient on the lagged dependent variable. Consequently, the results of including a lagged dependent variable may well be interpreted as supporting a partial adjustment model when in fact this is not the correct specification. As an example of this problem consider the following simple model. The equations

$$(1') \qquad \begin{aligned} y_t &= a + bz_t + u_t \\ u_t &= pu_{t-1} + e_t \qquad e_t \sim N(0, \sigma^2) \end{aligned}$$

could be written as

$$(2') \qquad y_t = a(1-p) + bz_t - bpz_{t-1} + py_{t-1} + e_t$$

(over)

There are two reasons why we think that it is unlikely that the error terms of these equations are serially correlated. First, the regressions were run again to include the dependent variable lagged two quarters. If serial correlation were present it is more than likely that the coefficient on the dependent variable lagged two periods would be significant (see fn. 19). In all instances this coefficient was insignificantly different from zero and the other coefficients were not significantly changed. Second, we have applied a test devised by J. Durbin which also suggests that there is no first-order serial correlation. This statistic, D, is asymptotically valid and will test for first-order positive serial correlation in the presence of lagged dependent variables. The statistic is defined as

$$(18) \qquad D = a\sqrt{\frac{n}{1 - n\hat{v}(b)}}$$

where $a=1-\frac{1}{2}d$, d is the Durbin-Watson statistic, n is the sample size, and $\hat{v}(b)$ the estimate of the variance of the coefficient of the lagged dependent variable. The statistic D is tested as a standard normal deviate so that if D is positive and greater than 1.96 then the null hypothesis is rejected and it is concluded that positive first-order serial correlation exists. It can be seen from Table 1 that D is uniformly less than 1.96 and therefore we conclude that there is no positive first-order serial correlation and therefore no serious bias with respect to the estimate of the adjustment speed.

It would appear from the results, and the discussion so far, that the concept of an effective price as a rationing device is a very powerful one. It is very successful at explaining the decisions of large groups of people to buy foreign or domestic goods. In order to get some feel for the quantitative importance of the variations of the ω elements of the effective price vector we conducted the following experiment. The relative actual price and time trend variables were set at their mean values and the observations on each of the excess demand variables were placed in the regression equations and the predicted X_2/X_1 ratios calculated. In all instances the Koyck lag was allowed to remain operative so that the predicted values are to be interpreted as short-run responses. The proportionate range of the predicted foreign-domestic goods ratio under different specifications of ω proved to be approximately 20 percent. Consequently, it would appear that the notion of an effec-

---

which, apart from the different error specification, is operationally equivalent to adding a partial adjustment response to the model

$$(3') \qquad y_t = a + bz_t - cz_{t-1} + v_t$$

to give

$$(4') \quad y_t = \lambda a + b\lambda z_t - c\lambda z_{t-1} + (1-\lambda)y_{t-1} + \lambda v_t$$

Griliches has suggested that these two models (2′) and (4′) may be distinguished by the following rule of thumb. If the coefficient on z_{t-1} is equal to minus the product of the coefficient of z_t and y_{t-1} then this can be taken as evidence that the model is not a partial adjustment model. If (2′) and (4′) are generalized to include more exogenous variables then it can be seen that the general form of these equations is very similar to the general form of the equations listed in Table 1. It is quite conceivable therefore that what we have interpreted to be a partial adjustment model is, in fact, the outcome of positive first-order serial correlation. For example, the coefficient on the lagged waiting time variable in regression 2 is not that different from minus the product of the coefficient on the current waiting time and the coefficient on the lagged dependent variable, that is $-(.57)(.68) = -.39 \simeq -.43$. Consequently our interpretation of the lagged dependent variable may be incorrect. There is, however, other evidence which supports the partial adjustment hypothesis.

A generalization of (2′) indicates that all exogenous variables, lagged one period, should enter into the estimated equation. In the context of the regressions reported in Table 1 this would suggest that both the lagged values of relative prices and the rate of change of prices should enter into the regressions with negative and significant coefficients. It can be seen from Table 1 that the restriction on the coefficient of the lagged rate of change of prices does not hold. Furthermore, in a number of unreported regressions it was found that when the lagged relative price term was included its coefficient was insignificantly different from zero.

tive price, which in the short run clears the market to a large extent by variations in its non-price component, is quantitatively an important concept.

The Rate of Capacity Utilization

In many of the discussions of the relationship between imports and the pressure of demand upon domestic resources, the rate of capacity utilization in the manufacturing sector is often explicitly mentioned as a causal factor explaining some of the variation in imports over the business cycle (see, for example, Branson). In this paper we have argued that it is not the capacity constraints per se that affect the allocation between domestic and foreign goods but rather the relative effective prices that face consumers. However, if the effective domestic price varied directly with the rate of capacity utilization of the domestic manufacturing sector then the rate of capacity utilization may serve as a simple proxy variable for the cyclical variations of the effective price.

In Table 2 we list some of the results obtained from hypothesizing that ω is a log-linear function of the rate of capacity utilization in the manufacturing sector. The capacity variable was subject to a number of transformations not all of which appear in Table 2. In different regressions it was lagged one, two, or three quarters and in order to test for non-linearities the variable was subdivided into observations when the capacity utilization rate was less than 81 percent (c_1), between 81 and 89 percent (c_2), and greater than 89 percent (c_3). We also tested the relationship between ω and the rate of capacity utilization for cyclical asymmetries by separating the capacity variable into one series of observations when the rate of capacity utilization was increasing ($c\uparrow$) and another series of observations when the rate of capacity utilization was decreasing ($c\downarrow$). We now turn to an analysis of these results.

Consider first regressions 1 to 3 of Table 2. These contain ω as a function of the rate of capacity utilization alone. The first overall impressions are that the R^2 are a little lower than in Table 1 and that the capacity variables do not perform particularly well. The coefficients of the capacity variables possess the correct signs but are significant at the 5 percent level only when the rate of capacity utilization is not subject to a subdivision into more than one series. There is no evidence of an asymmetrical response of the foreign domestic good ratio to variations of capacity utilization over the cycle (see equation (2)) nor is there any evidence that the relationship is non-linear (see equation (3)).

When the rate of capacity utilization is lagged it becomes increasingly insignificant and often exhibits the incorrect sign. For those variables which are also common to Table 1 the magnitude of their coefficients and the degree of significance of the estimates proved to be, in general, insensitive to the inclusion of the capacity utilization rate as a measure of the ω elements of the effective price. Consequently, we will not discuss these results in detail.

In regression equations 4, 5, and 6, the capacity utilization variable is combined with some of the variables suggested by the theory of Section II. In these circumstances the capacity utilization variable is usually insignificant although often marginally so. There is again no evidence of non-linearities or asymmetrical responses over the cycle.

These results suggest that little is to be gained from specifying ω as a function of the capacity utilization rate, or from introducing the rate of capacity utilization into the specification of the effective price. Furthermore, they add support to the view that it is not bottleneck and capacity constraints per se that affect the allocation of demand between foreign and

TABLE 2—THE RATIO OF IMPORTS TO DOMESTIC GOODS PRODUCTION REGRESSED AGAINST VARIOUS SPECIFICATIONS OF THE EFFECTIVE PRICE RATIO
Quarterly Data 1948–68 (t-values in parentheses)

Equations	1	2	3	4	5	6
Constant	−.3058 (.71)	−.2609 (.58)	−.4023 (.49)	.0823 (.20)	.1210 (2.84)	−.1385 (.33)
$\frac{P_1}{P_2}$	.3751 (2.89)	.3660 (2.78)	.3801 (2.90)	.9398 (5.17)	.5348 (3.99)	.5117 (3.72)
t	.0010 (2.05)	.0009 (2.00)	.0011 (2.36)	.0007 (1.80)	.0010 (2.32)	.0010 (2.34)
$\frac{X_{2t-1}}{X_{1t-1}}$	.7708 (10.85)	.7735 (10.74)	.7612 (10.69)	.6657 (10.12)	.6890 (9.55)	.7034 (9.49)
c_t	.1984 (2.00)			.0569 (.61)		.1989 (2.07)
$c\uparrow_t$		.1867 (1.80)			.1486 (1.53)	
$c\downarrow_t$		.1867 (1.78)			.1485 (1.51)	
c_{1t}			.2270 (1.21)			
c_{2t}			.2220 (1.21)			
c_{3t}			.2256 (1.25)			
l_{at}				.5490 (4.94)		
l_{at-1}				−.4196 (4.48)		
$\frac{P_{1t}}{P_{1t-1}}$					1.7551 (2.75)	1.7280 (2.45)
$\frac{P_{1t-1}}{P_{1t-2}}$					1.3250 (1.89)	
S.E.R.	.047	.046	.047	.041	.044	.046
R^2	.92	.92	.92	.94	.93	.92
D	1.50	1.51	1.20	1.24	.36	.74

Symbols: c_t = the rate of capacity utilization in the manufacturing sector; $c\uparrow_t$ = the rate of capacity utilization when the capacity utilization index is increasing; $c\downarrow_t$ = the rate of capacity utilization when the capacity utilization index is decreasing; c_{1t} = the rate of capacity utilization < 81 percent; c_{2t} = the rate of capacity utilization > 81 and < 89 percent; c_{3t} = the rate of capacity utilization > 89 perrent.

Source: The capacity utilization indices are taken from the *Federal Reserve Bulletin*. D is a measure of serial correlation. All variables are expressed in logarithms, seasonally adjusted, and measured in constant prices.

domestic goods, but relative effective prices.

Preliminary Application of the Model to Disaggregated Data

Any attempt to apply the model to data at lower levels of aggregation runs into quite serious data problems. It is difficult to match each category of imports with the relevant category of domestic goods and, in general, data series of the price indices of imports, and for most of the proxy variables suggested in Section II, are unavailable. However, despite these problems it may be worthwhile to report very briefly on some preliminary results.[20]

Table 3 lists the results of the first attempts to apply the model to lower levels of data aggregation. The categories were chosen to reduce the data problems, although they are considerable. Consider the first two categories, producer and consumer durables. The estimates of the magnitude and degree of significance of the coefficients of the relative price term, the lagged dependent variable, and the time trend are very similar to those of the aggregate data. The coefficient on the lagged dependent variable is still very significant and the response elasticity remains approximately .3. The coefficient on the time trend is again positive and significant. The short-run elasticity of substitution is significant and approximately .6. The proxy variables for the ω elements of the effective price possess the correct sign and are generally significant.

The least satisfactory results occurred in the third category, consumer nondurables. The price elasticity of substitution is insignificant, although it possesses the correct sign when the waiting time and its rate of change are added to the equation. Both waiting time variables possess the correct sign, but only one is significant at conventional levels.

As a generalization it could be said that the results support the hypotheses that the allocation of resources between domestic and foreign goods is responsive to variations in the non-price rationing variables discussed earlier. However, the model outlined in Section II does not perform quite as well at this level of aggregation. Whether this stems wholly from the inadequacy of the data (see Table 3 for the data description) or from the inadequacy of the model is impossible to say at this moment. There is obviously a need for considerably more work on the disaggregated data series.

IV. Conclusions

1. This paper is primarily concerned with two hypotheses. First, in the short run producers do not adjust prices to meet fluctuations of demand, but use other rationing devices, such as waiting times, credit terms, and even the enthusiasm with which they react to the enquiries of prospective customers. Second, the allocation of demand between domestic and foreign goods is responsive to these short-run rationing devices. A test of these hypotheses against the behavior of the ratio of *U.S.* imports to the production of domestic goods from the 1948–68 data reveals that the evidence supporting these hypotheses is quite considerable.

2. As a result of data difficulties these short-run rationing variables cannot be measured directly. The model of the firm outlined in Section II suggests that the quoted waiting time might be measured by the estimated delivery lag and its rate of change. Other proxy variables include the inventory level and its rate of change, and the rate of change of prices. When

[20] The theory of the previous sections can be applied to the disaggregated data by the extension of the utility tree to further sub-levels. Where the commodity in question is a producer good the utility function is replaced by a production function and the same analysis is applicable.

Table 3—Import Demand Functions for Various Categories of Imports
Quarterly Data 1953–68 (t-values in brackets)

Equations	Producer Durables 1	Producer Durables 2	Producer Durables 3	Producer Durables 4	Consumer Durables 1	Consumer Durables 2	Consumer Durables 3	Consumer Nondurables 1	Consumer Nondurables 2
Constant	−2.1928	−2.2307	−2.1353	−2.1800	−1.2368	−1.2180	−1.0239	−1.5795	−1.0561
	(3.92)	(4.25)	(4.02)	(4.27)	(3.47)	(3.94)	(2.85)	(3.01)	(1.90)
$\frac{P_{1t}}{P_{2t}}$	.6390	.7376	.6155	.6962	.6029	.6460	.5430	− .1083	.3324
	(2.87)	(3.49)	(2.88)	(3.35)	(2.42)	(2.30)	(2.23)	(.20)	(.62)
t	.0087	.0088	.0077	.0082	.0056	.0058	.0045	.0077	.0107
	(3.39)	(3.98)	(2.52)	(2.78)	(3.39)	(2.94)	(2.71)	(2.82)	(3.02)
$\frac{X_{2t-1}}{X_{1t-1}}$	.6374	.6337	.6574	.6467	.6755	.6737	.7295	.7168	.6483
	(7.02)	(7.44)	(7.03)	(7.20)	(7.36)	(7.29)	(7.91)	(7.63)	(6.28)
$\frac{P_{1t-1}}{P_{1t-2}}$		3.1733		2.5565			2.0455		
		(2.89)		(2.32)			(2.14)		
l_{at-1}						.2583			.4413
						(1.20)			(2.77)
l_{at-2}			.2967	.2433		− .2526			− .2109
			(2.51)	(2.10)		(1.31)			(1.32)
l_{at-3}			− .3267	− .2590					
			(2.77)	(2.22)					
S.E.R.	.059	.056	.056	.054	.056	.056	.055	.074	.070
R^2	.98	.99	.98	.99	.97	.97	.97	.97	.98
D	1.06	.25	1.06	.64	.72	1.20	.48	.96	.96

Symbols are defined in Table 1: Dependent Variables: Producer durables: Imports of capital equipment divided by Producers' durable equipment; Consumer durables: Imports of consumer durables, manufactured, divided by consumer expenditure on durables minus imports of consumer durables and consumer expenditure on automobiles and parts. Consumer nondurables: Imports of consumer nondurables, manufactured, divided by consumer expenditure on nondurable goods minus imports of consumer durables and consumption expenditure on food and beverages. Price Indices: Implicit price deflators of imports, producer durable equipment, personal consumption expenditure (durable goods), personal consumption expenditure (nondurable goods). Delivery lags: Calculated from unfilled orders, durable goods industries, and unfilled orders, nondurable goods industries. All data are taken from the *Survey of Current Business*. Data are seasonally adjusted, in constant prices and expressed as logarithms.

these variables are included in the demand function each possesses the predicted sign and magnitude and is statistically significant. A simple measure of the magnitude of domestic capacity constraints—the rate of utilization of capacity in the manufacturing sector—does not perform particularly well.

3. The inclusion of these proxy variables to measure variations in effective prices leads to an estimate of the price elasticity of substitution between foreign and domestic goods which is larger, and statistically more significant, than estimates derived from traditional specifications of demand equations. The estimates of the elasticity of substitution imply a short-run price elasticity for all imports of between .5 and unity and a long-run elasticity of demand of approximately three.

4. After allowance has been made for past changes in relative prices and the

non-price rationing variables, there has been a steady trend towards increasing the proportion of goods consumed which are imported.

5. Preliminary estimation of the model suggests that the above hypotheses may be supported when the data is further disaggregated but there remains considerable work to be done at lower levels of the data aggregation to derive conceptually adequate data and to test the model more fully.

6. In conclusion, the evidence presented here would suggest that on both theoretical and empirical levels more attention could be given to short-run disequilibrium situations and to variables which have not been traditionally included in demand and supply equations. Our model represents a step in this direction. In particular the theoretical model would benefit from a dynamic specification.

REFERENCES

R. J. Ball, J. R. Eaton, and M. D. Steuer, "The Relation between United Kingdom Export Performance in Manufactures and the Internal Pressure of Demand," *Econ. J.*, Sept. 1966, *76*, 501--19.

R. J. Ball and K. Marwah, "The U.S. Demand for Imports, 1948–58," *Rev. Econ. Statist.*, Nov. 1962, *44*, 395–401.

W. H. Branson, "A Disaggregated Model of the U.S. Balance of Trade," *Staff Economic Papers*, Board of Governors of the Federal Reserve System.

F. Brechling and J. N. Wolfe, "The End of Stop-Go," *Lloyds Bank. Rev.*, Jan. 1965, *73*, 23–30.

P. A. David and Th. Van De Klundert, "Biased Efficiency Growth and Capital Labour Substitution in the U.S., 1899–1960," *Amer. Econ. Rev.*, June 1965, *55*, 355–94.

J. Durbin, "Testing for Serial Correlation in Least Squares Regression when some of the Regressors are Lagged Dependent Variables," *Econometrica*, forthcoming.

O. Eckstein and G. Fromm, "The Price Equation," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1159–83.

J. Enzler, "Revised Indexes of Manufacturing Capacity and Capacity Utilization," *Fed. Res. Bull.*, July 1967, *53*, 1096–98.

W. Godley and J. Shepherd, "Forecasting Imports," *Nat. Inst. Econ. Rev.*, Aug. 1965, *33*, 35–42.

S. M. Goldman and H. Uzawa, "A Note on Separability in Demand Analysis," *Econometrica*, July 1964, *32*, 387–98.

Z. Griliches, "Distributed Lags: A Survey," *Econometrica*, Jan. 1967, *35*, 16–49.

K. Lancaster, "A New Approach to Consumer Theory," *J. Polit. Econ.*, Apr. 1966, *74*, 132–57.

W. Lederer, E. M. Parrish, and S. Pizer, "The Balance of International Payments: Fourth Quarter and Year 1965," *Surv. Curr. Bus.*, U.S. Department of Commerce, Mar. 1966, *46*, 16–28.

E. Malinvaud, *Statistical Methods of Econometrics*, Amsterdam 1966.

E. S. Phelps, "The New Microeconomics in Inflation and Employment Theory," *Amer. Econ. Rev. Proc.*, May 1969, *59*, 147–60.

R. Rhomberg and L. Boissonneault, "The Foreign Sector" in J. S. Duesenberry, ed., *Brookings Quarterly Econometric Model of the United States;* Chicago 1965, 375–408.

P. A. Samuelson, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.

K. Sato, "A Two Level Constant Elasticity of Substitution Production Function," *Rev. Econ. Stud.*, Apr. 1967, *34*, 201–18.

M. D. Steuer, R. J. Ball, and J. R. Eaton, "The Effect of Waiting Times on Foreign Orders for Machine Tools," *Economica*, Nov. 1966, *33*, 387–403.

L. C. Thurow, "A Fiscal Policy Model of the United States," *Surv. Curr. Bus.*, U.S. Department of Commerce, June 1969, *49*, 45–64.

Surv. Curr. Bus., U.S. Department of Commerce, various issues.

U.S. Congress Joint Economic Committee Subcommittee on Economic Statistics, *Hearings*, Jan 24, 1961, Washington 1961.

Cropsharing Tenancy in Agriculture: A Theoretical and Empirical Analysis

By P. K. BARDHAN AND T. N. SRINIVASAN*

Cropsharing tenancy is one of the earliest forms of production organization in agriculture. It is still a matter of considerable importance in peasant agriculture in many countries. While since Adam Smith much of the discussion[1] in the literature has been on the comparative efficiency of this with other forms of land institutions, it is less easy to find an analysis of the economic factors that may explain variations in the incidence of share-cropping in different areas. Needless to say, land institutions are shaped by diverse historical, political, and sociological factors peculiar to different regions; but an economist persists in believing that the "relations of production" ultimately reflect some of the basic economic "forces of production." The object of this paper is first to identify some of the relevant economic factors in terms of a simple theoretical framework and then check for their actual significance in an empirical analysis of interregional cross-section data pertaining to Indian agriculture.

In Section I, we discuss the basic theoretical model and the properties of equilibrium; in Section II we analyze the comparative static results with respect to variations in the wage rate, and in Section III, those with respect to parameters representing different kinds of technical progress; in Section IV we report the empirical results confirming some of the hypotheses in the theoretical sections; in Section V we comment on the problem of cost-sharing by landlords, and in the concluding Section VI, we briefly touch on the problem of uncertainty.

I. The Model and the Properties of Equilibrium

Let us take agricultural output to be a single homogeneous commodity whose production requires, for the time being, only land and labor. A landless person has the option to lease in some land and cultivate it with his labor or to work as wage-laborer on somebody else's farm or in non-agricultural occupations. In the former case he gets the output from the leased-in land after payment of a percentage share to the landlord and in the latter case, he receives a wage rate per unit of labor. We shall assume this wage rate to be given. The level of the agricultural wage rate depends on a whole host of factors outside the agricultural sector; for example, given other things, one expects the rural wage rate to be higher in West Bengal than in, say, Orissa, simply because by any index the former State is much more industrialized and urbanized. In a simplified model it seems better to assume that the wage is given as a parameter to the agricultural sector than to assume that it is completely determined by endogenous factors within the sector. So if our landless person consumes all of his income, his consumption is given by

* The authors are professors of economics at the Indian Statistical Institute, New Delhi. We have benefitted from suggestions for improvement by Amartya Sen and the editor and from the discussion in seminars at M.I.T., Delhi School of Economics and Vishwabharati at Santiniketan, where an earlier version of the paper was presented. All errors are, of course, ours alone.

[1] See a brief summary of it in D. Gale Johnson.

(1) $$C^1 = (1 - r)F(H, l_1) + l_2 w,$$

where r is the proportionate share of output paid as rent to the landlord, F is the production function, H is the amount of land hired by the tenant, l_1 is the amount of labor he devotes to the tenant farm, l_2 is the amount of labor he devotes to wage-paid occupations, and w is the given wage rate. We assume the production function is strictly concave and that its cross-partial derivatives are positive (i.e., the marginal product of one factor is an increasing function of the use of the other factor).

If we assume, without loss of generality, that the sharecropper has only one unit of labor available to him, then he devotes l_1 amount of time to his tenanted land, l_2 to outside work, and $(1-l_1-l_2)$ amount to leisure. If he is a maximizer, he maximizes a utility function

$$U^1(C^1, 1 - l_1 - l_2),$$

subject to the budget constraint in equation (1). We assume that the utility function is strictly concave and that neither consumption nor leisure is an inferior good.

If we confine ourselves to an interior maximum,[2] the necessary conditions are:

(2) $$(1 - r)F_1 = 0$$

(3) $$(1 - r)F_2 - w = 0$$

(4) $$U_1^1 w - U_2^1 = 0$$

As usual, subscript 1 refers to the first derivative with respect to the first argument and subscript 2 to the second argument.

[2] Empirically this is the more interesting case. There is plenty of evidence in India that tenants are also part-time agricultural laborers or that primarily agricultural labor households have income from tenant cultivation. On the landlord side also, simultaneous leasing-out of land and self-cultivation of the rest is quite frequent.

We might briefly comment on these conditions. Equation (2) implies that the sharecropper will tend to lease in more and more land until the marginal product of land is driven to zero.[3] Equation (3) refers to the usual result that a sharecropper will not use as much of a variable factor (in this case, labor) as an owner-farmer would; this leads to the standard case of inefficiency in sharecropping that economists from Adam Smith through Marshall have noted. Equation (4) also is a standard result: the optimum allocation of time between leisure and earning income implies that the ratio of their marginal utilities should be equalized to the wage rate.

The Jacobian matrix[4] $[a_{ij}]$ of equations (2), (3), and (4) will have the following elements:

$$a_{11} = (1 - r)F_{11}; \qquad a_{12} = (1 - r)F_{12};$$
$$a_{13} = a_{23} = a_{31} = 0; \quad a_{21} = (1 - r)F_{21};$$
$$a_{22} = (1 - r)F_{22};$$
$$a_{32} = a_{33} = \frac{U_2^1}{U_1^1}\left[U_{11}^1 \frac{U_2^1}{U_1^1} + U_{22}^1 \frac{U_1^1}{U_2^1} - 2U_{21}^1\right]$$

Totally differentiating (2), (3), and (4) with respect to r (with w given), we get

(5) $$[a_{ij}]\begin{bmatrix} \dfrac{dH}{dr} \\ \dfrac{dl_1}{dr} \\ \dfrac{dl_2}{dr} \end{bmatrix} = \begin{bmatrix} 0 \\ F_2 \\ F\left(U_{11}^1 \dfrac{U_2^1}{U_1^1} - U_{21}^1\right) \end{bmatrix}$$

[3] Particular attention to this condition was drawn by Johnson.

[4] The Jacobian $[a_{ij}]$ is negative, since with strictly concave production function, $F_{11}F_{22} > (F_{12})^2$, and with no inferior goods in the utility function, it is easy to show that $a_{33} < 0$.

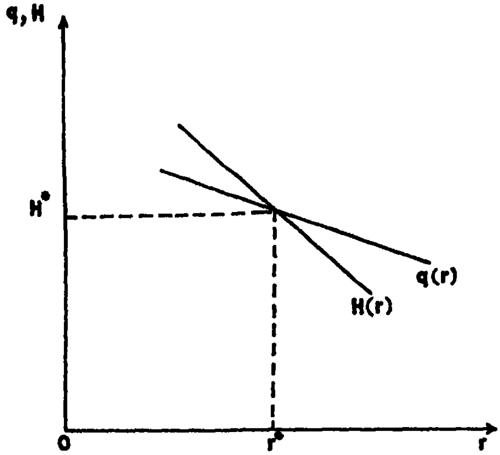


FIGURE 1. DEMAND AND SUPPLY IN THE MARKET FOR LAND LEASES

Using Cramer's Rule for solving simultaneous equations we get:

$$(6)\quad \frac{dH}{dr} = \frac{-F_2F_{12}}{(1-r)[F_{11}F_{22}-(F_{12})^2]} < 0$$

$$(7)\quad \frac{dl_1}{dr} = \frac{F_2F_{11}}{(1-r)[F_{11}F_{22}-(F_{12})^2]} < 0$$

Equations (6) and (7) imply that under our assumptions the amount of land leased in by the tenant as well as the amount of labor devoted to it, is a declining function of the rental share. Figure 1 depicts $H(r)$, the demand function for land to be leased in by sharecroppers.

Let us now look at the supply side. The land owner has the option to cultivate land with the use of own and hired labor or to lease out land to sharecroppers. In the former case, he has to pay hired labor at the given wage rate and in the latter, he gets only a share in the tenant's output. Assuming again that all of income is consumed, his consumption is

$$(8)\quad C^2 = G(1-q, x+y) - wx + rF(q, L)$$

where G is the production function in his own cultivation; assuming that the land owner has one unit of land, q is the proportion (or amount) of land he leases out to sharecroppers retaining $(1-q)$ for self-cultivation; x is the amount of hired labor and y is the amount of own labor (for simplicity we are assuming that these two types of labor are qualitatively similar), and L is the amount of labor the sharecropper puts in on the leased-out land.

If, again without loss of generality, we assume that the land owner has one unit of time available to him, then $(1-y)$ is the amount of time he takes out as leisure. He thus maximizes a utility function

$$U^2(C^2, 1-y)$$

subject to the budget constraint in equation (8). The properties of G and U^2 functions are assumed to be similar to those of F and U^1 functions we have encountered before. In the maximization process we should note that the land owner can vary q, x, and y, but he cannot choose L; it is up to the sharecropper to decide how much labor he would like to use on the tenanted land.

Once again confining ourselves to an interior maximum the necessary conditions are:

$$(9)\quad G_1 - rF_1 = 0$$

$$(10)\quad G_2 - w = 0$$

$$(11)\quad U_1^2G_2 - U_2^2 = 0$$

Equation (9) implies that the marginal product of land in self-cultivation should be equal to the share of the marginal product of land that land owner receives from leased-out land. Equation (10) implies that the marginal product of labor in self-cultivation should be equal to the wage rate. Equation (11) implies that the optimum allocation of time between leisure and earning income implies an equality of the ratio of their marginal utilities to the marginal productivity of labor.

The Jacobian matrix[5] $[b_{ij}]$ of equations (9), (10), and (11) will have the following elements:

$$b_{11} = G_{11} + rF_{11}; \quad b_{12} = b_{13} = G_{12};$$
$$b_{21} = G_{21}; \quad b_{22} = b_{23} = G_{22};$$
$$b_{31} = U_1^2 \cdot G_{21}; \quad b_{32} = U_1^2 \cdot G_{22};$$
$$b_{33} = U_1^2 \cdot G_{22} + \frac{U_2^2}{U_1^2}\left[U_{11}^2 \frac{U_2^2}{U_1^2} + U_{22}^2 \frac{U_1^2}{U_2^2} - 2U_{21}^2\right]$$

Totally differentiating (9), (10), and (11) with respect to r (with w and L given), we get

$$(12) \quad [b_{ij}]\begin{bmatrix} \frac{dq}{dr} \\ \frac{dx}{dr} \\ \frac{dy}{dr} \end{bmatrix} = \begin{bmatrix} F_1 \\ 0 \\ F\left(U_{11}^2 \frac{U_2^2}{U_1} - U_{21}^2\right) \end{bmatrix}$$

Using Cramer's Rule,

$$(13) \quad \left.\frac{dq}{dr}\right|_{L=L_0} = \frac{-F_1 \cdot G_{22}}{[G_{11}G_{22} - (G_{12})^2 + rF_{11} \cdot G_{22}]}$$

Similarly,

$$(14) \quad \left.\frac{dq}{dL}\right|_{r=r_0} = \frac{-rF_{12} \cdot G_{22}}{[G_{11} \cdot G_{22} - (G_{12})^2 + rF_{11} \cdot G_{22}]} > 0$$

Taking the demand and supply side (for land-leases)[6] together, we can now write the equilibrium conditions[7] as

$$(15) \qquad H(r) = q(r, L)$$

$$(16) \qquad L = l_1(r)$$

Equations (15) and (16) may be rewritten as one condition:

$$(17) \qquad H(r) = q(r, l_1(r))$$

From (6) we know that $dH/dr < 0$ and from (7), (13), and (14) we can now write that around equilibrium

$$(18) \quad \frac{dq}{dr} = \left.\frac{dq}{dr}\right|_{L=L_0} + \left.\frac{dq}{dL}\right|_{r=r_0}\left(\frac{dl_1}{dr}\right) = \left.\frac{dq}{dL}\right|_{r=r_0}\left(\frac{dl_1}{dr}\right) < 0,$$

since the equilibrium $F_1 = 0$.

So around equilibrium the supply of land leased out is also a *declining* function of the rental share. But it can be easily checked from (6), (7), and (14) that around equilibrium

$$(19) \qquad \frac{dH}{dr} - \frac{dq}{dr} < 0$$

so that in Figure 1, the $H(r)$ curve must intersect the $q(r)$ curve from above. Using the standard tools of Walrasian stability analysis the equilibrium is unique and stable. H^* is the equilibrium amount (or percentage) of land leased out to sharecroppers and r^* is the competitively determined rental share.[8]

[5] Again the Jacobian in this case is negative with strictly concave production function and no inferior goods in the utility function.

[6] In this paper we ignore the market for land buying and selling and concentrate our attention on land renting.

[7] Note that in equilibrium x is not equal to l_2; otherwise the agricultural wage rate would have been endogenously determined.

[8] We have so far ignored any form of land lease at a given rental rate per unit of land (rather than at a given rental share of output). Johnson has suggested that where there is this form of renting simultaneously with sharecropping the equilibrium conditions are different from those under sharecropping alone. We shall, however, show here that both these forms of land-lease cannot coexist in our model.

(over)

Before we pass on to our comparative-static analysis, let us briefly comment on some points raised in a recent paper on sharecropping by S.N.S. Cheung (1968). His paper, like the present one, tries to change the analysis of sharecropping from its usual partial-equilibrium framework to a simple general equilibrium footing, although the number of choices available to the economic agents are larger in the present paper (for example, Cheung does not give landlords the option to self-cultivate). Some of the implications of our model seem to be in contradiction with those of Cheung. One of his major conclusions is that the prevailing impression of inefficient allocation of resources under sharecropping is wrong. This does not seem to tally with our equation (3) where, clearly, the sharecropper stops short of equalizing the marginal product of labor to the wage rate. The difference lies in the kind of maximization process Cheung carried out in Section III, pp. 1113–14 of his paper. There he maximizes only from the landlord's point of view,[9] whereas in this paper we determine the demand side from maximization by the tenant, just as the supply side is determined from landlord's maximizing decision. Further, Cheung's landlords can freely choose the amount of labor the sharecropper would devote to the tenant farm whereas in our paper this is decided by the sharecropper alone. It seems to us that unless the share contract explicitly contains stipulations regarding the intensity of use of labor by the sharecropper, the better assumption is to leave the decision to the sharecropper.

It is also difficult to understand why Cheung's landlord maximizes with respect to r, the rental share. Cheung seems to think that in the traditional analysis of sharecropping, "the writers fail to realize that the percentage shares and area rented under share tenancy are not mysteriously fixed but are competitively determined in the market"; yet in his analysis the landlord himself decides about the level of r. In a competitive situation one would have thought that each atomistic landlord and each atomistic tenant would take r as given, and out of their aggregate maximizing decisions r is determined in the market, as in our paper. It seems that only when the landlord has monopoly power in the market for land-leases can he choose the level of r.

Cheung also introduces an extra decision variable, viz. the number of parcels into which the landlord can subdivide his land leasing out each to a different tenant. In

---

Suppose (1) is rewritten as

$$C^1 = (1-r)F^1(H_1, l_1) + F^2(H - H_1, l_2) - R(H - H_1) + w(1 - l_1 - l_2),$$

where H_1 and l_1 are the amounts of land and labor, respectively, under sharecropping, F^1 is the production function on the sharecropped farm, F^2 is the production function on the land leased under the other form of renting, R is the given rental rate per unit of such land; and for simplification, let us ignore leisure. One of the necessary conditions for maximizing C^1 is

$$rF_1^1 \lesseqgtr 0 \quad \text{as } H_1 = \begin{cases} 0 \\ \epsilon\ (0, H) \\ H \end{cases}$$

Similarly, landlord's consumption is

$$C^2 = G(1-q, 1+x) - wx + rF^1(q_1, L_1) + R(q - q_1)$$

where q_1 is the amount of land leased out under sharecropping and once again we ignore leisure. One of the necessary conditions of maximizing C^2 is

$$rF_1^1 - R \lesseqgtr 0 \quad \text{as } q_1 = \begin{cases} 0 \\ \epsilon\ (0, q) \\ q \end{cases}$$

Now putting these two necessary conditions together we can now say that $H_1 = q_1 \epsilon (0, q)$ leads to a contradiction, for then $F_1^1 = 0$ which means $R = 0$ which means no land will be rented out under the alternative form of renting.

So it seems in our model all of leased land is either under sharecropping or under a given rental rate per unit of land. We have chosen in this paper to devote our analysis only to the former problem.

[9] The only way the tenant's decision enters in any way in Cheung's model is that the tenant's income from cultivation should be equal to the amount he would have earned had he been a wage laborer, i.e., in our notation, $wl_1 = (1-r)F$. We have seen above that if the tenant can devote part of his labor to tenant-cultivation and part to wage-employment, this changes to $w = (1-r)F_2$.

that case there is no reason why he should ignore a similar decision variable on the part of the tenant, i.e., the tenant can also choose the number of landlords from each of whom he leases in a parcel of land.

In our model above we have ignored these two decision variables. Let us briefly mention here how one may incorporate them in our model. If n is the number of landlords from whom the tenant leases in a parcel of land, equation (1) has to be rewritten as:

$$(20)\qquad C^1 = n(1-r)F(h, l_1) + l_2 w,$$

where h is the size of each tenanted parcel and l_1 is the amount of labor devoted to it. The utility function has to be rewritten as

$$U^1(C_1, 1 - l_1 n - l_2)$$

When this function is maximized with respect to h, l_1, and l_2, we get exactly the same necessary conditions as (2), (3), and (4). It is to be noted particularly that the inefficiency in labor use under sharecropping that was implied by (3) remains valid even in this case.

Differentiating U^1 with respect to n, we get

$$(21)\qquad \frac{\partial U^1}{\partial n} = (1-r)U_1^1[F - F_2 \cdot l_1] > 0$$

This implies that the tenant continuously gains from parcellization. This is expected because of the diminishing returns to scale implied by our strictly-concave production function. For any meaningful analysis one should impose an upper bound on the number of feasible parcels, say $\bar{n}$, and if we assume that in our model n will always be equal to $\bar{n}$.

Similarly, from the landlord side also it is easy to show that a strictly-concave production function implies

$$\frac{\partial U^2}{\partial_m} > 0,$$

when m is the number of parcels into which the landlord can subdivide his land leasing out each to a different tenant. Once again we shall assume an upper bound, $\bar{m}$, which will be reached.

In the rest of the analysis we shall, therefore, again ignore the problem of parcellization.

II. Parametric Shifts in the Wage Rate

In the model above we have taken the wage rate as exogenously given to the agricultural sector. Let us now find out the implications of parametric shifts in the wage rate.

Differentiating with respect to w in equations (2), (3), and (4), keeping r fixed for the moment, we get

$$(22)\qquad [a_{ij}]\begin{bmatrix} H_w \\ l_{1w} \\ l_{2w} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ -U_1^1 - l_2\left(U_{11}^1 \dfrac{U_2^1}{U_1} - U_{21}^1\right) \end{bmatrix}$$

where

$$H_w = \left.\frac{dH}{dw}\right|_{r=r_0}, \qquad l_{1w} = \left.\frac{dl_1}{dw}\right|_{r=r_0}, \text{ etc.}$$

Using Cramer's Rule again,

$$(23)\qquad H_w = \frac{-F_{12}}{(1-r)[F_{11}\cdot F_{22} - (F_{12})^2]} < 0$$

$$(24)\qquad l_{1w} = \frac{F_{11}}{(1-r)[F_{11}\cdot F_{22} - (F_{12})^2]} < 0$$

Equations (23) and (24) imply that for a given rental share, the higher the wage rate the lower is the tenant's demand for leasing-in land (since wage employment is more attractive now) and also the amount of labor devoted to tenanted land.

Differentiating with respect to w in

equations (9), (10), and (11), keeping r and L fixed for the moment, we get

$$(25)\quad [b_{ij}]\begin{bmatrix} q_w \\ x_w \\ y_w \end{bmatrix} = \begin{bmatrix} 0 \\ x\left(U_{11}^2 \dfrac{U_2^2}{U_1^2} - U_{21}^2\right) \end{bmatrix}$$

We may solve for

$$(26)\quad q_w = \frac{G_{12}}{[G_{11}\cdot G_{22}-(G_{12})^2+rF_{11}\cdot G_{22}]} > 0$$

This means that given r and L, a rise in the wage rate for hired labor induces landowners to lease out more land.

But in equilibrium with a rise in the wage rate r cannot remain fixed. Let us rewrite equilibrium condition (17) as

$$(27)\quad H(r, w) = q(r, w, l_1(r, w))$$

Using (14), (19), (23), (24), and (26) we can get equation (28), which means that a parametric rise in the wage rate brings down the equilibrium rental share.

Let us now find out how such a parametric shift in the wage rate affects the equilibrium amount (or percentage) of land leased out under cropsharing.

From (27), and using (6), (23), and (28), we obtain (29) which implies that a parametric rise in the wage rate leads to a *rise* in the equilibrium amount (and percentage) of land leased out under cropsharing. In our empirical analysis of cross-sectional data, we shall accordingly try to test the hypothesis that a region with a higher wage rate will tend to have a larger incidence of cropsharing tenancy.

III. Technical Progress

In all of our analysis above we assumed that output depends on the use of labor and land and our production functions did not admit of any technical change. We shall now introduce parameters representing different kinds of technical progress.

Let us rewrite tenant's consumption equation (1) as

$$(30)\quad C^1 = (1 - r)\rho F(\mu H, \lambda l_1) + l_2 w$$

where ρ, μ, and λ are multiplicative technical progress parameters. We shall take three types of technical progress corresponding to types familiar in the growth literature:

a) Land-augmenting technical progress, when ρ and λ are constant but μ shifts;

b) Hicks-Neutral technical progress, when μ and λ are constant but ρ shifts and

c) Labor-augmenting technical pro-

$$(28)\quad \frac{dr}{dw} = \left\{(1 - r)[F_{11}\cdot F_{22} - (F_{12})^2][G_{11}\cdot G_{22} - (G_{12})^2 + rF_{11}\cdot G_{22}]\left[\frac{\partial H}{\partial r} - \frac{\partial q}{\partial r}\right]\right\}^{-1}$$
$$\cdot [F_{12}\{G_{11}\cdot G_{22} - (G_{12})^2\} + (1 - r)G_{12}\{F_{11}\cdot F_{22} - (F_{12})^2\}] < 0$$

$$(29)\quad \frac{dH}{dw} = H_w + \frac{\partial H}{\partial r}\cdot\frac{dr}{dw}$$
$$= -\left\{(1 - r)[F_{11}\cdot F_{22} - (F_{12})^2][G_{11}\cdot G_{22} - (G_{12})^2 + rF_{11}\cdot G_{22}]\right.$$
$$\left.\cdot\left[\frac{\partial H}{\partial r} - \frac{\partial q}{\partial r}\right]\right\}^{-1} F_2\cdot F_{12}\cdot G_{12} > 0$$

gress, when ρ and μ are constant but λ shifts. We shall assume that the same parameters obtain with respect to G, the landlord's self-cultivation function.

Taking all these parameters into account, the necessary conditions of tenant's utility maximization given by equations (2), (3), and (4) will have to be rewritten as

$$(31) \qquad (1-r)\rho\mu\cdot F_1 = 0$$

$$(32) \qquad (1-r)\rho\lambda F_2 - w = 0$$

$$(33) \qquad U_1^1 w - U_2^1 = 0$$

Similarly, the landlord's utility maximization will involve

$$(34) \qquad \mu\rho(G_1 - rF_1) = 0$$

$$(35) \qquad \lambda\rho G_2 - w = 0$$

$$(36) \qquad \lambda\rho U_1^2 G_2 - U_2^2 = 0$$

We shall now use these equations for our comparative-static analysis.

Let us first take case a), i.e., land-augmenting technical progress, and assume for simplicity that $\rho=\lambda=1$. Using differentiation and Cramer's Rule as before it is easy to show from equations (31), (32), and (33) that

$$(37) \qquad H_\mu = \left.\frac{dH}{d\mu}\right|_{r=r_0} = -\frac{H}{\mu} < 0$$

and

$$(38) \qquad l_{1\mu} = \left.\frac{dl_1}{d\mu}\right|_{r=r_0} = 0$$

Similarly, from equations (34), (35), and (36),

$$(39) \qquad q_\mu = \left.\frac{dq}{d\mu}\right|_{r=r_0} = \{\mu[G_{11}G_{22} - (G_{12})^2 + rF_{11}G_{22}]\}^{-1} \cdot [(1-q)\{G_{11}G_{22} - (G_{12})^2\} - qrF_{11}\cdot G_{22}]$$

$$(40) \qquad \frac{\partial q}{\partial L} = -rF_{12}\cdot G_{22} \cdot \{\mu[G_{11}G_{22} - (G_{12})^2 + rF_{11}G_{22}]\}^{-1} > 0$$

The equilibrium condition (17) is now rewritten as

$$(41) \qquad H(r,\mu) = q(r,\mu,l_1(r))$$

Using (19), (37), (39), and (41) we can get

$$(42) \qquad \frac{dr}{d\mu} = \left\{\mu[G_{11}G_{22} - (G_{12})^2 + rF_{11}G_{22}] \cdot \left[\frac{\partial H}{\partial r} - \frac{\partial q}{\partial r}\right]\right\}^{-1} \cdot [G_{11}G_{22} - (G_{12})^2] < 0$$

In other words, the rental share is a declining function of the land-augmenting technical progress parameter.

Using (6), (7), (19), (37), (39), (40), and (42), we obtain equation (43) which shows that with land-augmenting technical progress the equilibrium amount (and percentage) of land leased out under cropsharing goes *up*.

Let us now take case b), i.e., neutral technical progress and assume for simplicity that $\mu=\lambda=1$. From (31), (32), and (33),

$$(43) \qquad \frac{dH}{d\mu} = -\left\{\mu^2(1-r)[F_{11}F_{22} - (F_{12})^2][G_{11}G_{22} - (G_{12})^2 + rF_{11}G_{22}] \cdot \left[\frac{\partial H}{\partial r} - \frac{\partial q}{\partial r}\right]\right\}^{-1} [G_{11}G_{22} - (G_{12})^2](1-q) > 0$$

$$H_\rho = \frac{dH}{d\rho}\bigg|_{r=r_0} = \frac{F_2 F_{12}}{\rho[F_{11}F_{22} - (F_{12})^2]} > 0 \tag{44}$$

$$l_{1\rho} = \frac{dl_1}{d\rho}\bigg|_{r=r_0} = -\frac{F_2 \cdot F_{11}}{\rho[F_{11}F_{22} - (F_{12})^2]} > 0 \tag{45}$$

Similarly, from (34) and (35), and (36),

$$q_\rho = \frac{dq}{d\rho}\bigg|_{r=r_0} = \frac{-G_2 \cdot G_{12}}{\rho[G_{11}G_{22} - (G_{12})^2 + rF_{11}G_{22}]} < 0 \tag{46}$$

The equilibrium condition (17) is now rewritten as

$$H(r, \rho) = q(r, \rho, l_1(r, \rho)) \tag{47}$$

From (14), (19), (44), (45), (46), and (47), we obtain equation (48) which shows that with a rise in the neutral technical progress parameter the rental share also goes up.

Using (6), (7), (19), (44), (45), and (46), we obtain equation (49) which shows that with neutral technical progress the equilibrium amount (and percentage) of land leased out under cropsharing goes *down*.

Let us finally take case c), i.e., labor-augmenting technical progress and assume for simplicity that $\rho=\mu=1$. From (31), (32), and (33)

$$H_\lambda = \frac{dH}{d\lambda}\bigg|_{r=r_0} = \frac{F_2 F_{12}}{\lambda[F_{11} \cdot F_{22} - (F_{12})^2]} > 0 \tag{50}$$

$$l_{1\lambda} = \frac{dl_1}{d\lambda}\bigg|_{r=r_0} = \frac{-[\lambda l_1\{F_{11}F_{22} - (F_{12})^2\} + F_2 F_{11}]}{\lambda^2[F_{11}F_{22} - (F_{12})^2]} \tag{51}$$

Similarly, from (34), (35), and (36),

$$q_\lambda = \frac{dq}{d\lambda}\bigg|_{r=r_0} = \frac{-[G_2 G_{12} + r\lambda L F_{12} \cdot G_{22}]}{\lambda[G_{11}G_{22} - (G_{12})^2 + rF_{11}G_{22}]} \tag{52}$$

$$\frac{\partial q}{\partial L} = \frac{-r\lambda F_{12} G_{22}}{[G_{11}G_{22} - (G_{12})^2 + rF_{11}G_{22}]} > 0 \tag{53}$$

The equilibrium condition (17) is now rewritten as

$$H(r, \lambda) = q(r, \lambda, l_1(r, \lambda)) \tag{54}$$

Using (19), (50), (51), (52), (53), and (54), one can show that $dr/d\lambda$ is exactly equal to the expression on the right-hand side of

---

$$\frac{dr}{d\rho} = -\left\{\rho[F_{11}F_{22} - (F_{12})^2][G_{11}G_{22} - (G_{12})^2 + rF_{11}G_{22}]\left[\frac{\partial H}{\partial r} - \frac{\partial q}{\partial r}\right]\right\}^{-1} \cdot [G_2 G_{12}\{F_{11}F_{22} - (F_{12})^2\} + F_2 F_{12}\{G_{11}G_{22} - (G_{12})^2\}] > 0 \tag{48}$$

$$\frac{dH}{d\rho} = \left\{\rho[F_{11} \cdot F_{22} - (F_{12})^2](1 - r)[G_{11}G_{22} - (G_{12})^2 + rF_{11}G_{22}] \cdot \left[\frac{\partial H}{\partial r} - \frac{\partial q}{\partial r}\right]\right\}^{-1} F_2 \cdot F_{12} G_2 \cdot G_{12} < 0 \tag{49}$$

(48) with λ substituted for ρ. So we can say that with rise in the labor-augmenting technical progress parameter the rental share goes up.

Using (6), (7), (19), (50), (51), (52), (53), and (54), we can also show that $dH/d\lambda$ is exactly equal to the expression on the right-hand side of (49) with λ substituted for ρ. So we can say that with labor-augmenting technical progress the equilibrium amount (and percentage) of land leased out under cropsharing goes *down*.

We might mention here two examples of how the results in this section may be useful towards explaining observed phenomena. The first is in terms of *U.S.* agriculture and the second in terms of Indian agriculture. It is well known that sharecropping was quite prevalent in the Mississippi Delta at least up to the 1930's. It is sometimes maintained[10] that since then large-scale mechanization substituting for hand-picking of corn and cotton has contributed to a rapid decline in sharecropping. *If* the introduction of labor-saving technical equipments is regarded as a case (instead of factor substitution along the same isoquant) of labor-augmenting technical progress (i.e., of an increase in the effective supply of labor), then this phenomenon in the southern United States agriculture is consistent with our conclusion above that labor-augmenting technical progress leads to a fall in the incidence of sharecropping.

In Indian agriculture, irrigation may largely be regarded as a factor bringing about land-augmenting technical progress. By improving land-productivity and facilitating double cropping, irrigation serves to increase the effective supply of land: in efficiency units a piece of irrigated land may be regarded as a multiple of a piece of unirrigated land of same acreage. In terms of the conclusion of our crude model regarding land-augmenting technical progress we then expect that with an increase in the importance of irrigation, given other things, the incidence of sharecropping should *increase*. In terms of a cross-sectional study, this means that we expect a larger percentage of area to be under sharecropping in better irrigated regions. In our empirical analysis with Indian data we have tried to test this hypothesis.

IV. Empirical Analysis

On the basis of our comparative-static analysis in Sections II and III, we expect a positive correlation between the importance of cropsharing tenancy in India and the agricultural wage rate or irrigation. In this section, we try to test it with the help of cross-sectional data across States and across some villages in India. Our dependent variable is the percentage of operated area leased in under cropsharing tenancy (we shall call it S), which represents an index of the incidence of sharecropping. Our main independent variables are w, the agricultural wage rate, and I, the percentage of operated area irrigated. Needless to say, in the real world S will depend on various other economic and particularly noneconomic factors. But our purpose is only to test if we have been able to identify at least two major economic factors explaining variations in the incidence of sharecropping and if the signs of the regression coefficients for these two variables are as expected from our comparative-static analysis. Since we do not claim to have identified all the major explanatory variables, we are not unduly perturbed if our R^2 in the subsequent analysis is not always very high. Besides, tenancy data, particularly in situations where there is a lot of land reform legislation at least in the statute books, may be

[10] See, for example, Richard Day.

quite treacherous sometimes, and even extremely good statistical fits to such data may be suspect. Regarding the tenancy figures we shall be using we may indicate here our general belief that they are more likely to be underestimates of the actual incidence of tenancy than otherwise; we proceed on the inevitable assumption that this under-reporting bias is fairly uniformly distributed across regions.

Let us first report the results on the State-level data for the year 1960–61. On the basis of National Sample Survey data for thirteen States, we had the following least-squares estimate for a log-linear equation:

$$\text{(55)}\quad \log S = \underset{(9.9)}{2.967} + \underset{(0.64)}{1.428}\log w^{**} + \underset{(0.23)}{0.574}\log I^{**} - \underset{(2.4)}{2.663}\log B; \quad R^{2**} = .602$$

Here and later we use ** for significant at 5 percent level and * for significant at 1 percent level.

In (55) the regression coefficients for w and I are significant (at 5 percent level) and of expected positive signs. Let us explain the variable B. In all of our analysis above we assumed that the market for land-lease is competitive. But the Indian land market is very much imperfect. In equation (55) we have taken B as a very crude index for the bargaining power of land-owners: It is the Lorenz concentration index of land ownership in the rural areas of each State. The assumption is that the larger is this concentration index the stronger is the bargaining power of the landlords in the land market.

What sign do we expect for the regression coefficient of B? Without going through detailed analysis we may just invoke a familiar result of the theory of monopoly. A monopolist tends to restrict sales of output below the competitive level. Similarly, one may expect that a landlord having a strong bargaining power in the land market will restrict leasing out below the level that a competitive market will reach. On this ground we expect, other things remaining the same, a negative correlation between S and B. In (55) the coefficient for B is not very significant (it is significant only at 30 percent level), but the sign *is* negative.

The multiple correlation coefficient, R, for the estimate as a whole is significant at 5 percent level.

Next we report results on the basis of forty villages covered by farm management surveys of the Government of India, Ministry of Food and Agriculture: ten in Ferozepur (Punjab), ten in Amritsar (Punjab), ten in Sambalpur (Orissa), and ten in West Godavari (Andhra Pradesh). The least-square estimate in this case is

$$\text{(56)}\quad \log S = -\underset{(0.914)}{0.603} + \underset{(0.710)}{1.452}\log w^{**} + \underset{(0.254)}{0.691}\log I^{*}; \quad R^{2*} = .35$$

Once again coeficients for w and I are quite significant and of expected positive signs. For lack of data we could not use the variable B in this case. Although R^2 is not very high for the estimate, the multiple correlation coefficient is significant at 1 percent level.

We have also used, though somewhat less successfully, the data for thirty-nine villages in Punjab and Western U.P. covered by surveys of the Delhi University Agro-Economic Research Centre. The least-square estimate that gives the best fit is

$$\text{(57)}\quad S = \underset{(18.0)}{6.609} + \underset{(2.5)}{9.804}\, w^{*} + \underset{(0.066)}{0.026}\, I - \underset{(18.0)}{9.369}\, B; \quad R^{2*} = .386$$

The coefficients for w, I, and B (concentration index of land ownership) are all of expected signs. But it is significant only for w. The multiple correlation coefficient is significant at 1 percent level.

V. Cost-Sharing by Landlords

In this section we comment on the problem of cost-sharing by landlords in the context of cropsharing. Let us suppose that apart from land and labor there is a third input to production, fertilizers. The landlord may or may not share in the cost incurred by the tenant on fertilizers. We shall discuss the equilibrium relationship between crop share and cost share.

Let the tenant's production function be $F(H, N, l)$ where H and l denote land and labor inputs and N the input of fertilizer. We shall assume that F is strictly concave and exhibits diminishing returns to scale. Let p be the price per unit of N in terms of output. Let r denote the landlord's share in output and β his share in fertilizer cost. Let w be the externally given wage rate that the tenant can earn as a wage laborer. Let the total labor that the tenant can render be unity (for simplification, we ignore leisure). Then the tenant's income (consumption) is:

$$C^1 = (1 - r)F(H, N, l) + w(1 - l) - (1 - \beta)pN \tag{58}$$

Maximizing C^1 with respect to the tenant's choice variables H, N, and l we get (for an interior maximum)

$$(1 - r)F_1 = 0 \tag{59}$$

$$(1 - r)F_2 = (1 - \beta)p \tag{60}$$

$$(1 - r)F_3 = w \tag{61}$$

where F_i denotes the partial derivative of F with respect to its ith argument.

Let us denote by $|J|$ the determinant of the Jacobian matrix

$$J = \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{bmatrix}$$

where F_{ij} denotes the partial derivative of F_i with respect to its jth argument. By assumption $F_{ij}=F_{ji}$ and concavity of F implies $|J|<0$, $F_{ii}<0$ $i=1, 2, 3$ and the determinants

$$\begin{vmatrix} F_{ii} & F_{ij} \\ F_{ji} & F_{jj} \end{vmatrix} > 0 \quad \text{for } i, j = 1, 2, 3, \; i \neq j$$

Let us further assume that $F_{ij} \geq 0$ for $i \neq j$. In other words, let us suppose that the marginal product of any input is a nondecreasing function of each of the other two inputs.

Using (59)–(61) it is easy to derive (62)–(67). These equations yield the ex-

$$H_r \equiv \frac{\partial H}{\partial r} = \frac{1}{(1 - r)^2 |J|}[(F_{13}F_{32} - F_{12}F_{33})(1 - \alpha)p + (F_{12}F_{23} - F_{13}F_{22})w] < 0 \tag{62}$$

$$N_r \equiv \frac{\partial N}{\partial r} = \frac{1}{(1 - r)^2 |J|}[(F_{11}F_{33} - F_{13}^2)(1 - \alpha)p + (F_{21}F_{13} - F_{11}F_{23})w] < 0 \tag{63}$$

$$l_r \equiv \frac{\partial l}{\partial r} = \frac{1}{(1 - r)^2 |J|}[(F_{12}F_{31} - F_{11}F_{32})(1 - \alpha)p + (F_{11}F_{22} - F_{12}^2)w] < 0 \tag{64}$$

$$H_\beta \equiv \frac{\partial H}{\partial \beta} = \frac{-p(F_{13}F_{32} - F_{12}F_{33})}{(1 - r)|J|} > 0 \tag{65}$$

$$N_\beta \equiv \frac{\partial N}{\partial \beta} = \frac{-p(F_{11}F_{33} - F_{13}^2)}{(1 - r)|J|} > 0 \tag{66}$$

$$l_\beta \equiv \frac{\partial l}{\partial \beta} = \frac{-p(F_{12}F_{31} - F_{11}F_{32})}{(1 - r)|J|} > 0 \tag{67}$$

pected results that (i) given β, the landlord's share in fertilizer costs (as well as w and p), the tenant will rent *less* land, put in *less* labor and buy *less* fertilizers the *larger* the value of r, the landlord's share in output, and (ii) given r, the tenant will rent *more* land, put in *more* labor and buy more fertilizers the *larger* the value of β.

Let us now turn to the landlord. Let us assume that he has one unit of land which he can either cultivate with hired labor or rent out to a sharecropper or do both in any desired proportion. Let q denote the proportion rented out. Let $G(1-q, x, y)$ denote the production function applicable to the land cultivated by the landlord with the help of hired labor, where x and y, respectively, denote the quantities of fertilizer and hired labor used. G is assumed to have the same properties as F. Let us assume that the landlord treats as given and independent of his own behavior the amount of fertilizers $\overline{N}$ and labor L put in by the tenant. With these assumptions, the landlord's income (consumption reduces) to:

$$(68)\qquad \begin{aligned} C^2 &= G(1-q, x, y) - px - wy \\ &\quad + rF(q, \overline{N}, L) - \beta p\overline{N} \end{aligned}$$

Maximizing C^2 with respect to choice variables q, x, and y we get (for an interior maximum):

$$(69)\qquad -G_1 + rF_1 = 0$$

$$(70)\qquad G_2 - p = 0$$

$$(71)\qquad G_3 - w = 0$$

Let us denote by $|J'|$ the determinant of the Jacobian matrix

$$J' = \begin{bmatrix} G_{11} + rF_{11} & -G_{12} & -G_{13} \\ -G_{21} & G_{22} & G_{23} \\ -G_{31} & G_{32} & G_{33} \end{bmatrix}$$

Concavity of F and G implies $|J'| < 0$,

$$\begin{vmatrix} G_{ii} & G_{ij} \\ G_{ji} & G_{jj} \end{vmatrix} > 0 \quad \text{for } i \neq j,\ F_{ii},\ G_{ii} < 0$$

As in the case of F, $G_{ij} > 0$ for $i \neq j$. It is easy to show that

$$(72)\qquad q_r \equiv \frac{\partial q}{\partial r} = \frac{-F_1(G_{22}G_{33} - G_{23}^2)}{|J'|} \geq 0$$

$$(73)\qquad q_N \equiv \frac{\partial q}{\partial \overline{N}} = \frac{-rF_{12}(G_{22}G_{33} - G_{23}^2)}{|J'|} \geq 0$$

$$(74)\qquad q_L \equiv \frac{\partial q}{\partial L} = \frac{-rF_{13}(G_{22}G_{33} - G_{23}^2)}{|J'|} \geq 0$$

It can be verified that x_r and y_r are nonpositive and $x_{\overline{N}}$, $y_{\overline{N}}$, x_L, y_L are all nonnegative. These are expected results: The landlord will rent out more land if his share in output of tenanted land is higher or if the tenant increase his inputs of fertilizers and labor. Also in each of these situations, the amount of land cultivated by the landlord himself will be correspondingly less. So will be the amounts of labor hired by him and the fertilizers purchased.

It should be noted that the landlord's choice variables do not depend directly on β, his share in the cost of fertilizers used by the tenant. This follows from our assumption that the landlord behaves as if he has no influence over the amounts of labor and fertilizer inputs used by the tenant.

Let us now turn to the equilibrium conditions:

$$(75)\qquad H = q$$

$$(76)\qquad N = \overline{N}$$

$$(77)\qquad l = L$$

We have thus nine equations, (59)–(61), (69)–(71), and (75)–(77) to determine ten unknowns, H, N, l, q, $\overline{N}$, L, x, y, r, and β. Thus we have one unknown more than the number of equations and except in special cases, there will be more than one solution to the unknowns. In principle one can

$$(78) \qquad \frac{dr}{d\beta} = \frac{(1-r)p(F_{13}F_{32} - F_{12}F_{33})}{[(F_{13}F_{32} - F_{12}F_{33})(1-\beta)p + (F_{12}F_{23} - F_{13}F_{22})w]} = \frac{-H_\beta}{H_r} > 0$$

solve the system for nine of the unknowns as functions of the tenth. Mathematics apart, there is one natural way of looking at the problem; i.e., as a market for land leases with the landlords as suppliers and tenants as demanders. There are two price-like variables; r the landlord's share of output and β the landlord's share in fertilizer cost. However, a single market-clearing condition that amount of land rented out by the landlords equals that leased in by the tenants can only determine one of these price variables as a function of the other. We treat r as a function of β and examine the behavior of r as β varies.

It can be shown that:

$$\frac{dr}{d\beta} = -\frac{q_l \cdot l_\beta + q_N N_\beta - H_\beta}{q_l \cdot l_r + q_N N_r - H_r + q_r}$$

After substituting the values of q_l, q_N, etc. and noting that $q_r = 0$ in equilibrium, we get equation (78) which shows that $dr/d\beta$ is positive. This fact implies that the larger is the share of costs borne by landlords, the larger is the equilibrium rental share of crop. This rather straightforward result is worth keeping in mind, particularly because in the standard discussion on cost-sharing (see, for example, Adams and Rask) it is ignored by the assumption of keeping the rental share fixed and varying the cost share. Our result provides a simple explanation of a common observation in Indian agriculture: when the landlord participates in the costs, the rental share he is paid is much higher than otherwise. For example, the *Farm Management Survey in West Bengal* points to the *Bhagchasi* system in which the landlord does not share in the costs and gets about half of the crop produced by the sharecropper whereas in the alternative *Krishani* system, the landlord himself covers most of the non-labor costs and usually gets about two-thirds of the crop raised by the sharecropper. Our result also implies that a government cannot try to implement rent-regulating legislation and at the same time also expect to induce the landlords to share more in tenant's fertilizer costs through sheer exhortations.

Equation (78) has other interesting implications. It can be rewritten as $H_r dr/d_\beta + H_\beta = dH/d\beta = 0$. This means that regardless of the value of β in the relevant range (0, 1), the equilibrium share of the landlord in output will so adjust that the equilibrium amount of land leased out by the landlord (and leased in by the tenant) remains invariant.

Let us now consider a special case of the above model in which the tenant does not render wage labor and devotes all his working hours to the sharecropped land. For this case it can be shown that $dr/d\beta = 1 - r/1 - \beta$ which can be integrated to yield $r = (1-\theta) + \theta\beta$ where θ is some constant. From the fact that even if cost-sharing were absent (i.e., $\beta = 0$), the equilibrium cropshare r will be positive, it follows that $0 < \theta < 1$. It also follows that $r > \beta$ for all β except $\beta = 1$ when $r = 1$. However $r = 1$ rules out any tenancy. Hence we can assert that for all relevant values of β, i.e., in the interval (0, 1), the equilibrium output share of the landlord *exceeds* his share in costs.[11]

It can also be shown for this special case (and for it only), that $dC^1/d\beta < 0$,

[11] We could prove this result only for the special case. In the general case it follows from (78) that $dr/d\beta < (1-r)/(1-\beta)$ and hence $r < (1-\theta) + \theta\beta$, with $0 < \theta < 1$. However this does not preclude $r \leq \beta$.

which implies that if the tenant had the choice of β, he would choose β to be zero. In other words, the tenant would prefer the landlord *not sharing* in costs. This apparently paradoxical result is explained as follows: we have seen that $dr/d\beta > 0$ implying that the larger the values of β the larger is the equilibrium value of r. Apparently the effect of lower r more than offsets the effect of lower β so that the optimum value of β from the tenant's point of view is zero. It can be shown that the landlord would prefer $\beta = 1$ so that $r = 1$. But this would rule out tenancy altogether. If, however, the social welfare indicator is the sum of the consumption of landlord and tenant, this sum can be shown to be invariant with respect to β so that the optimum β from the social point of view is any β in the interval (0, 1).

VI. Uncertainty and Other Matters

In all of the analyses above we have ignored the intimate relationship of crop-sharing arrangements with problems of coping with uncertainty. In this paper we do not intend to deal with this important but very difficult issue in any detail. Let us just indicate here a possible theoretical way of approaching it, although we do not have many results to report.

Let us suppose A is a parameter representing production uncertainty (say, due to fluctuations of weather) and suppose equation (1) of Section I is rewritten (ignoring leisure) as

$$C^1 = (1 - r)AF(H, l) + w(1 - l) \tag{79}$$

and equation (8) as

$$C^2 = AG(1-q, 1+x) - wx + rAF(q, L) \tag{80}$$

Let us take a simple characterization of the uncertainty parameter A such that

$$A = \alpha u + \beta \tag{81}$$

where u is the random variate.

With E as the expectation operator and V for variance,

$$E(A) = \alpha E(u) + \beta \quad \text{and} \quad V(A) = \alpha^2 V(u)$$

For capturing the effects of parametric shifts in uncertainty we can vary the variance of A keeping the mean value of A constant by varying α and β in such a way that $d\beta/d\alpha = -E(u)$. Thus

$$\frac{dA}{d\alpha} = u + \frac{d\beta}{d\alpha} = u - E(u) \tag{82}$$

It is to be noted that with $\alpha > 0$, $(u - E(u))$ has the same sign as $(A - E(A))$.

We shall assume that both the tenant and the landlord maximize the expected value of their respective utility functions.

The necessary conditions for tenant's interior maximization are:

$$F_1 = 0 \tag{83}$$

$$E[U^{1\prime}(C^1)\{(1 - r)AF_2 - w\}] = 0 \tag{84}$$

Similarly, the necessary conditions for the landlord's interior maximum are:

$$rF_1 - G_1 = 0 \tag{85}$$

$$E[U^{2\prime}(C^2)\{AG_2 - w\}] = 0 \tag{86}$$

Let us now analyze the equations (83)–(86). The Jacobian matrix $[a_{ij}]$ of equations (83) and (84) has the following elements:

$$a_{11} = F_{11};\ a_{12} = F_{12};$$
$$a_{21} = E[U^{1\prime}(C^1)(1 - r)AF_{21}]$$
$$a_{22} = E[U^{1\prime}(C^1)(1 - r)AF_{22}] + E[U^{1\prime\prime}(C^1)\{(1 - r)AF_2 - w\}]$$

It is easy to check that the Jacobian is positive. From (82), (83), and (84), we may derive equation (87). The non-zero element in the column vector on the right-hand side of (87) can be shown to be positive[12] if $e^1 = -U^{1\prime\prime}(C^1) \cdot C^1/U^{1\prime}(C_1)$, the

[12] Under the assumption of constant e^1

$$\phi(A) = [U^{1\prime}(C^1)F_2 + U^{1\prime\prime}(C^1)\{(1 - r)AF_2 - w\}F]$$

$$(87)\quad [a_{ij}]\begin{bmatrix}\frac{dH}{d\alpha}\\ \frac{dl}{d\alpha}\end{bmatrix}=\begin{bmatrix}0\\ -(1-r)E[\{U^{1\prime}(C^1)F_2+U^{1\prime\prime}(C^1)((1-r)AF_2-w)F\}(u-Eu)]\end{bmatrix}$$

$$(88)\quad [b_{ij}]\begin{bmatrix}\frac{dq}{d\alpha}\\ \frac{dx}{d\alpha}\end{bmatrix}=\begin{bmatrix}0\\ -E[\{U^{2\prime}(C^2)G_2+U^{2\prime\prime}(C^2)(AG_2-w)(G+rF)\}(u-E(u))]\end{bmatrix}$$

index of relative risk aversion, is assumed to be constant. This means that $dH/d\alpha<0$; in other words, other things remaining the same, a larger importance of production uncertainty implies that the tenant will lease in *less* land under cropsharing.

Similarly, on the landlord side let us take equations (85) and (86). The Jacobian matrix $[b_{ij}]$ has the following elements.

$$b_{11}=G_{11}+rF_{11};\ b_{12}=-G_{12};$$
$$b_{21}=-E[U^{2\prime}(C^2)AG_{21}];$$
$$b_{22}=E[U^{2\prime}(C^2)AG_{22}]+E[U^{2\prime\prime}(C^2)+\{AG_2-w\}^2]$$

It is easy to check that the Jacobian in this case also is positive. From (85) and (86), we obtain equation (88). Using the same technique as in footnote 12 we can prove that the non-zero element in the column vector on the right-hand side of (88) is positive if $e^2=-U^{2\prime\prime}(C^2)\cdot C^2/U^{2\prime}(C^2)$, the index of relative risk aversion on the part of landlords, is assumed to be constant. This means that $dq/d\alpha>0$; in other words, given other things (particularly r, w, and L), a larger importance of production uncertainty implies that the landlord will tend to lease out *more* land to sharecroppers.

Because of the complicated calculations involved we have not been able to go any further. In any case, it seems that in our model increase in production uncertainty induces the tenant and the landlord to go in opposite directions in the market for land leases. We have not been able to find out the direction in the movement of the *equilibrium* proportion of land under sharecropping. Obviously much more work needs to be done in this area[13] before even elementary comparative-static propositions can be made.

We may also point out here that apart from production uncertainty, other forms of uncertainty will also seriously affect the equilibrium percentage of area under sharecropping. For example, in our analysis above, increased production uncertainty induces the tenant to devote less labor to the tenanted farm and more to wage labor (this is largely because work

is a declining function of A. Hence

and $\quad \phi(A)>\phi(EA)\quad$ when $A<EA$

$\quad\quad \phi(A)<\phi(EA)\quad$ when $A>EA$

so

$$E\phi(A)(-EA)<\phi(EA)\cdot E(A-EA)=0$$

[13] For not very rigorous treatments of this problem, see Cheung and Rao. Among other things, Cheung expects a larger incidence of share-tenancy in areas with higher production uncertainty and he mentions evidence from Chinese agriculture. In India, irrigation, apart from facilitating technical progress, has also a protective role against bad crop weather (there is some evidence of a negative correlation between percentage of area irrigated and variance of output) and yet in our empirical analysis there is a *positive* correlation between importance of irrigation and that of share-tenancy.

outside brings remuneration in the form of certain wage income whereas work on leased-in land involves uncertain income); but in a more realistic analysis, uncertainty of getting employment outside often drives the tenant to prefer a cropsharing arrangement through which he shares the production uncertainty with the landlord. Similarly, on the landlord side there are the uncertainties of possible default of rent by the tenant in bad years and of availability of wage labor in peak seasons.

To sum up, in this paper we have tried to identify some of the economic factors that may contribute towards explaining regional variations in the incidence of cropsharing tenancy and to test for their significance on the basis of Indian State-level and village-level data. In the process we have built a simple theoretical model for analyzing sharecropping which may be useful in any further discussion on this form of tenancy. We have also tried to throw some theoretical light on the problem of cost-sharing by landlords.

Needless to say, our model here is very crude and it fails to capture many of the economic and institutional aspects of cropsharing tenancy as it is observed in peasant economies. In particular, we have hardly scratched the surface of any satisfactory analysis of various kinds of uncertainties that are relevant to cropsharing arrangements. We have also ignored the varying degrees of coexistence of cropsharing with other forms of tenancy, like fixed rent, on account of differences in uncertainty, the problems connected with the duration of lease contracts, and the various ways in which land-market imperfections (landlord dominance on the one hand and protective tenancy and rent-regulating legislation along with ways of evading them on the other) distort the simple rules of the competitive game we have assumed in most of this paper. Yet we believe that our basic framework will be useful in further explorations on this subject.

REFERENCES

D. W. Adams and N. Rask, "Economics of Cost-Sharing Leases in Less Developed Countries," *Amer. J. Agr. Econ.*, Nov. 1968, *50*, 935–42.

S. N. S. Cheung, "Private Property Rights and Share-Cropping," *J. Polit. Econ.*, Nov.-Dec. 1968, *76*, 1107–22.

———, "Transaction Costs, Risk Aversion and the Choice of Contractual Arrangements," *J. Law Econ.*, Apr. 1969, *12*, 23–42.

R. H. Day, "The Economics of Technological Change and the Demise of the Share-Cropper," *Amer. Econ. Rev.*, June 1967, *57*, 427–49.

D. G. Johnson, "Resource Allocation Under Share Contracts," *J. Polit. Econ.*, Apr. 1950, *58*, 111–23.

C. H. H. Rao, "Entrepreneurship, Management, and Farm Tenure Systems," unpublished.

Government of India, Ministry of Food and Agriculture, *Farm Management Survey in West Bengal.*

On the Theory of the Competitive Firm Under Price Uncertainty

By AGNAR SANDMO*

In recent years several contributions have been made to the theory of the firm under uncertainty, removing the assumption that the demand for the product is known with certainty at the time when the output decision is made. In most of these papers the assumption is made that the objective of the firm is to maximize expected profits.[1] This is hardly a very satisfactory assumption, since it completely rules out risk averse behavior, and so many elementary facts of economic life seem to indicate a prevalence of risk aversion.

The present paper is intended as a systematic study of the theory of the competitive firm under price uncertainty and risk aversion. We assume that the decision on the volume of output to be produced must be taken prior to the sales date, at which the market price becomes known. The firm's beliefs about the sales price can be summarized in a subjective probability distribution. However, since the firm is unable to influence this distribution, the basic assumption that the firm is a price taker is retained—in a probabilistic sense.[2]

It is perhaps most natural to interpret the model of the paper as being concerned with the short run. The firm makes its output decisions with sole regard for short-run profits and does not consider the relationship between this output policy and long-run policies for investment and finance. In a sense, it is a weakness of the model that it takes no account of this interrelatedness; but it may also be considered a strength, because a more complete model would make it necessary to draw up a much larger and more detailed list of assumptions about the economic environment of the firm than is needed for the present paper. The results presented here are thus compatible with several alternative sets of assumptions about investment opportunities, financial markets, and the structure of ownership. It is only essential to assume that short-run output decisions are dominated by a concern for short-run profits.

Occasionally, especially in Section III, we shall also find it convenient to use the model to analyze some long-run problems. It then becomes necessary to assume that these long-run elements have implicitly been accounted for. This is hardly satisfactory. Still, it is a useful simplification with long traditions in the theory of the firm.

We shall assume that the firm's attitude towards risk can be summarized by a von Neumann-Morgenstern utility function. This may be a strong assumption, because in many firms decisions are typically taken by a group of individuals, and group preferences may not always satisfy the

* Professor of economics, Norwegian School of Economics and Business Administration. This paper was written while I was a fellow of the Center for Operations Research and Econometrics, Université Catholique de Louvain. I am indebted to Jacques Drèze and Jean Jaskold Gabszewicz for their valuable comments.

[1] For some examples see the papers by Drèze and Gabszewicz, Kenneth Smith, Edward Zabel, and the book by Clement Tisdell.

[2] A similar approach is taken by Phoebus Dhrymes, Saul Hymans, John McCall, Bernt Stigum (1969a) and Hayne Leland (1969). Some interesting comments can also be found in Karl Borch (ch. 12, especially pp. 171–73).

transitivity axiom required for the existence of a utility function. It is therefore possible that this approach implicitly assumes that the firm's reactions to changes in its environment are more predictable and stable than they really are. However, there are still many firms in which decisions are essentially made by one person, and there are presumably firms in which preferences are sufficiently similar within the group of decision makers to guarantee the existence of a group preference function. This provides justification for the approach taken in this paper.

I. Optimal Output under Uncertainty

We assume that the objective of the firm is to maximize the expected utility of profits. The utility function of the firm is a concave, continuous and differentiable function of profits, so that

$$U'(\pi) > 0, \quad U''(\pi) < 0 \tag{1}$$

Thus, the firm is assumed to be risk averse. It is well known that in order for a utility function to satisfy the von Neumann-Morgenstern axioms without giving rise to St. Petersburg phenomena, it must be bounded from above.[3] Strictly speaking, then, equation (1) holds only in the range below the upper bound of U.

The cost function of the firm is

$$F(x) = C(x) + B, \tag{2}$$

where x is output, $C(x)$ is the variable cost function, and B is "fixed cost." About the variable cost function we make the following general assumptions:

$$C(0) = 0, \quad C'(x) > 0 \tag{3}$$

The firm's profit function can now be defined as

$$\pi(x) = px - C(x) - B, \tag{4}$$

where p is the price of output, assumed to be a (subjectively) random variable with density function $f(p)$ and expected value $E[p]=\mu$. Naturally, p is restricted to be nonnegative. This means that, once x has been chosen, the firm's maximum loss is $(-C(x)-B)$. Clearly also, $\pi(0)=-B$.

The expected utility of profits can be written as

$$E[U(px - C(x) - B)],$$

where E is the expectations operator. Differentiating with respect to x, we obtain as necessary and sufficient conditions for a maximum:

$$E[U'(\pi)(p - C'(x))] = 0, \tag{5}$$

$$D = E[U''(\pi)(p - C'(x))^2 - U'(\pi)C''(x)] < 0 \tag{6}$$

It is interesting to note that in order for the second-order condition (6) to hold, it is not necessary to assume increasing marginal cost.

For the remainder of Section I and in Section II, we assume that (5) and (6) determine a non-zero, finite and unique solution to the maximization problem. The problems of existence and of corner solutions will be discussed in Section III.

One question which is naturally raised by the introduction of price uncertainty is this: how does the optimal output compare with the well-known competitive solution under certainty? Under certainty, the solution is characterized by equality between price and marginal cost. There is no obvious way of making such a comparison, but one possible and appealing specification of the problem is this: what is the optimal output under uncertainty as compared with the situation where the price is known to be equal to the expected value of the original distribution? Referring to the latter level of output as the certainty output, we shall now show that *under price uncertainty, output is smaller than the*

[3] See on this point Kenneth Arrow, who also argues that U must be bounded from below.

certainty output. This is a generalization of a theorem of McCall, who proves a similar result for the case of a utility function with constant absolute risk aversion.

The first-order condition (5) can be written as

$$E[U'(\pi)p] = E[U'(\pi)C'(x)] \tag{7}$$

Subtract $E[U'(\pi)\mu]$ on each side of this equation. We then get

$$E[U'(\pi)(p-\mu)] = E[U'(\pi)(C'(x)-\mu)] \tag{8}$$

Since $E[\pi]=\mu x - C(x) - B$ (from the definition of profits), we have that $\pi = E[\pi] + (p-\mu)x$. Clearly

$$U'(\pi) \leqq U'(E[\pi]) \quad \text{if } p \geqq \mu \tag{9}$$

It follows immediately that

$$U'(\pi)(p-\mu) \leqq U'(E[\pi])(p-\mu) \tag{10}$$

This inequality holds for all p. For if $p \leqq \mu$, the inequality sign in (9) is reversed, but then multiplication by $(p-\mu)$ will still make $\leqq$ hold in (10). Taking expectations on both sides of (10) and noting that $U'(E[\pi])$ is a given number, we obtain

$$E[U'(\pi)(p-\mu)] \leqq U'(E[\pi])E[p-\mu]$$

But, here the right-hand side is equal to zero by definition, and so the left-hand side is negative. Then we know that the right-hand side of (8) is negative also. But this can be written as

$$E[U'(\pi)](C'(x)-\mu) \leqq 0,$$

and, since marginal utility is always positive, this implies

$$C'(x) \leqq \mu \tag{11}$$

That is, optimal output is characterized by marginal cost being less than the expected price. Now under certainty the only types of cost curves compatible with competitive assumptions are those for which the marginal cost curve is either everywhere increasing or else U-shaped. In those cases, (11) proves our statement above. Equation (11) is, of course, also valid for constant or decreasing marginal cost, but then the competitive output is not well defined.

This result is not the only conceivable answer to the question of the effect of uncertainty on the output decision. Following Jacques Drèze and Franco Modigliani, we may describe our result as concerned with the *overall* impact of uncertainty. However, one may also be interested in the question of the *marginal* impact; i.e., the effect of making a given distribution "slightly more risky." It is not obvious how this can be formalized; in the following we shall adopt a procedure used in Sandmo.

Let us define a small increase in risk as a "stretching" of the probability distribution around a constant mean. This requires the introduction of two shift parameters, one multiplicative and one additive. Thus, let us write price as

$$\gamma p + \theta,$$

where γ is the multiplicative shift parameter and θ is the additive one. An increase of γ alone (from the point $\gamma=1$, $\theta=0$) will "blow up" all values of p; it will therefore increase the mean as well as the variance. To restore the mean we have to reduce θ simultaneously, so that

$dE[\gamma p + \theta] = 0$, or $\mu d\gamma + d\theta = 0$, i.e.,

$$\frac{d\theta}{d\gamma} = -\mu \tag{12}$$

We can now write the profit function as $\pi(x) = (\gamma p + \theta)x - C(x) - B$ and differentiate with respect to γ, taking account of (12). The result is then

$$\frac{\partial x}{\partial \gamma} = -x \cdot \frac{1}{D} E[U''(\pi)(p-\mu)(p-C'(x))] - \frac{1}{D} E[U'(\pi)(p-\mu)] \tag{13}$$

Of these two terms, the last one is clearly

negative (from the proof above and from the second-order condition). However, the sign of the first term is in general indeterminate, so that at the present level of generality it does not seem possible to make a precise statement about the marginal impact of uncertainty.

There is one special case in which we would expect the marginal impact of uncertainty to become identical to the overall impact. That is in the case where we start from the certainty of $p=\mu$ and replace this certain price by a probability distribution with all outcomes concentrated in the neighborhood of μ. This is not too easily handled, since our stretching procedure breaks down in that case. However, we can get around this difficulty by noting that, when price is known to be equal to μ, we must have $C'(x)=\mu$. Then the first term in (13) becomes

$$-x \cdot \frac{1}{D} E[U''(\pi)(p-\mu)^2],$$

which is certainly negative. Thus, both terms in (13) are negative, and their signs depend only on the assumption of risk aversion. The connection with the overall impact of uncertainty is thereby established.

II. The Comparative Statics of the Firm

Simply assuming the existence of risk aversion is a very weak restriction on the firm's attitudes to risk. Further restrictions on the utility function may be introduced by means of the Arrow-Pratt risk aversion functions:

$$\text{Absolute risk aversion: } R_A(\pi) = -\frac{U''(\pi)}{U'(\pi)}$$

$$\text{Relative risk aversion: } R_R(\pi) = -\frac{U''(\pi)\pi}{U'(\pi)}$$

It seems reasonable to assume that $R_A(\pi)$ is a decreasing function of π. This would reflect the hypothesis that as a decision maker becomes wealthier (in terms of income, profit etc.), his risk premium for any risky prospect, defined as the difference between the mathematical expectation of the return from the prospect and its certainty equivalent, should decrease, or at least not increase. If $R_R(\pi)$ is increasing, this means that the elasticity of the risk premium with respect to π is less than one in absolute value. Arrow argues that there are good theoretical and empirical reasons for making this assumption, but the evidence for it does not seem conclusive, and we shall not commit ourselves to a specific hypothesis as to the form of $R_R(\pi)$.[4]

One of the basic results in the theory of the firm under certainty is that fixed costs do not matter in the sense that once a strictly positive output level has been chosen, this output is unaffected by an infinitesimal increase in fixed costs. This is not so under uncertainty. Differentiating in (5) with respect to B, we obtain

$$(14) \quad \frac{\partial x}{\partial B} = \frac{1}{D} E[U''(\pi)(p - C'(x))]$$

Decreasing absolute risk aversion is a necessary and sufficient condition for $\partial x/\partial B$ to be negative. The proof of this is as follows: Let $\bar{\pi}$ be the level of profits when $p=C'(x)$. Then, since $R_A(\pi)$ is decreasing[5]

[4] Some remarks on the empirical evidence can be found in the article by Joseph Stiglitz. For derivations of the risk aversion functions the reader is referred to the contributions of Arrow and John Pratt. Hypotheses about the risk aversion functions have been applied to portfolio theory by Arrow, to insurance purchasing and to taxation and risk-taking by Jan Mossin (1968a, b), and to the analysis of saving decisions by Sandmo. Several other examples of application could easily be given.

[5] This must be interpreted with care. We are interested in the properties of the risk aversion function at the optimum position, i.e., for the output level $x=x^*$ which is the solution to (5). For this given output level, (15) is certainly true. It is important to note that this *local* relationship is independent of the *global* lack of any one-to-one relationship between the algebraic signs of profits and marginal revenue.

(15) $R_A(\pi) \leqq R_A(\bar{\pi}) \quad \text{for } p - C'(x) \geqq 0$

Substituting from the definition of $R_A(\pi)$, we obtain

$$(16) \quad -\frac{U''(\pi)}{U'(\pi)} \leqq R_A(\bar{\pi}) \quad \text{for } p - C'(x) \geqq 0$$

(Note that $R_A(\bar{\pi})$ is a given number and not a random variable.) We know of course that

$$(17) \quad -U'(\pi)(p - C'(x)) \leqq 0 \text{ for } p - C'(x) \geqq 0,$$

since marginal utility is positive. Now multiply (16) by the left-hand side of (17). We then get

$$U''(\pi)(p - C'(x)) \geqq -R_A(\bar{\pi})U'(\pi)(p - C'(x))$$

This holds for all p. For if $p \leqq C'(x)$, the inequality in (16) is reversed, but so is that in (17). Now taking expected values we obtain

$$E[U''(\pi)(p - C'(x))] \geqq - R_A(\bar{\pi})E[U'(\pi)(p - C'(x))]$$

But by the first-order condition (5), the right-hand side is equal to zero, and the left-hand side is accordingly positive. But then the derivative (14) is negative and our proposition is proved.

Is this conclusion in itself intuitively plausible? This question may perhaps best be judged by considering whether a lump sum tax or a lump sum subsidy would be the most appropriate policy measure for making the firm increase its output. Economic intuition seems strongly to suggest the latter alternative, which is exactly what our result implies.

We turn now to an examination of the firm's supply function. Since the price is seen by the firm as a random variable, it does not make sense to speak about the effect of an "increase in price." It seems natural, however, to discuss the closely related problem of an increase in the mathematical expectation of the price with higher central moments constant. We can do this in the following way: Let us write price as $p+\theta$, where θ is again an additive shift parameter. Increasing θ is equivalent to moving the probability distribution to the right without changing its shape. Differentiating (5) with respect to θ and evaluating the derivative at $\theta=0$ we obtain

$$\frac{\partial x}{\partial \theta} = -x \cdot \frac{1}{D} E[U''(p - C'(x))] - \frac{1}{D} E[U'(\pi)],$$

or, substituting from (14),

$$(18) \quad \frac{\partial x}{\partial \theta} = -x\frac{\partial x}{\partial B} - \frac{1}{D} E[U'(\pi)]$$

This expression is similar to the Slutsky equation familiar from demand analysis. It says that the firm's response to an increase in expected price can be decomposed into two separate effects, one of which is analogous to a decrease in fixed costs, and the other one is a pure substitution effect. Of the latter effect we can immediately say that it is positive. As for the sign of the former effect we can draw on our previous result to conclude that *decreasing absolute risk aversion is a sufficient condition for $\partial x/\partial \theta$ to be positive,* i.e., for an upward-sloping supply curve. Again the implication of decreasing absolute risk aversion seems intuitively plausible. It implies, e.g., that in order to increase output the government should consider a per unit subsidy, rather than a per unit tax, as the appropriate policy measure.[6]

Another well-established result in the theory of the firm is that a change in a proportional rate of profit taxation will have no effect on the level of output. A priori there is no reason to expect this result to hold under uncertainty.

[6] The interested reader who wishes to see an example where the possibility of a downward-sloping supply curve does occur may consider the simple case of a quadratic utility function and constant marginal cost, where the supply curve bends backward for expected price sufficiently high.

With price uncertainty the question of loss offset provisions becomes important. If there is no loss offset, the profit function of the firm becomes

$$\pi(x) = \begin{cases} px - C(x) - B & \text{for } p \leqq \dfrac{C(x) + B}{x} \\ (px - C(x) - B)(1 - t) & \text{for } p > \dfrac{C(x) + B}{x} \end{cases}$$

On the other hand, if there is full loss offset, the profit function can be written as

$$\pi(x) = (px - C(x) - B)(1 - t) \quad \text{for all } p$$

It is not easy to decide which of these two assumptions is the more interesting and realistic one. Full loss offset presupposes that the firm or its owner(s) has other income from which any loss can be deducted. In fact, tax laws in many countries do provide for loss offset, either against other income or against future profits, so that there may be reasons for concentrating attention on this case.[7]

With full loss offset expected utility is

$$E[U((px - C(x) - B)(1 - t))],$$

and the first-order condition becomes

$$E[U'(\pi)(p - C'(x))] = 0, \tag{19}$$

as before, since the multiplicative factor $(1-t)$ can be factored out.

Differentiating in (19) with respect to t yields

$$\frac{\partial x}{\partial t} = \frac{1}{1 - t} \cdot \frac{1}{D} \cdot E[U''(\pi)\pi(p - C'(x))] \tag{20}$$

It can be shown that *increasing the tax rate will increase, leave constant or reduce output according as relative risk aversion is increasing, constant, or decreasing.*

If $R_R(\pi)$ is increasing, we must have that

$$-\frac{U''(\pi)\pi}{U'(\pi)} \geqq R_R(\bar{\pi}) \quad \text{for } p - C'(x) \geqq 0 \tag{21}$$

Multiplying this by $-U'(\pi)(p - C'(x))$ yields

$$U''(\pi)\pi(p - C'(x)) \leqq -R_R(\bar{\pi})U'(\pi)(p - C'(x)),$$

and by the argument used in the proof above, this inequality holds for all p. Taking expectations, the right-hand side vanishes, and we have that

$$E[U''(\pi)\pi(p - C'(x))] \leqq 0$$

From this it follows that $\partial x/\partial t$ is positive in the case of increasing relative risk aversion. The proof of the rest of the statement follows immediately.

III. Profits, Entry, and Returns to Scale

It is well known that under certainty increasing marginal cost is necessary for the existence of a competitive optimum for the firm. This is not so under uncertainty, as we shall now demonstrate.[8]

Consider first the case where marginal cost is constant. Then concavity and boundedness of U as a function of π is sufficient to show that there exists a finite $x = x^*$ which gives a maximum of U. The case $C''(x) > 0$ is equally simple, because increasing marginal cost only reinforces the concavity of U as a function of x. It follows also that the case of a U-shaped marginal cost curve is only slightly more complicated· for then U will be concave in x in the region for which $C'(x) \geqq \min C'(x)$.

Note also that in the case of decreasing MC followed by constant MC the above

[7] This argument is not entirely satisfactory, however. If "other income" of "future profits" are at least partially determined by the firm's own actions, they should presumably be integrated into the model.

[8] For a rigorous discussion of the existence of optimal policies under uncertainty the reader is referred to Leland (1970).

argument remains valid; there will be a determinate optimal level of output for the firm. The troublesome case is where *MC* is everywhere decreasing and boundedness of the utility function no longer guarantees the existence of an optimal policy. However, it remains true that decreasing *MC* is not a sufficient condition for the nonexistence of an optimal output level; thus a market *may* be competitive even under this assumption.

So far, we have assumed the existence of an interior maximum for the firm; i.e., we have assumed that the optimal level of output is strictly positive. But we know from received theory that even if the condition "price=marginal cost" determines a local maximum of profits, the maximum need not, even if it is a unique interior maximum, give us the global maximum. The reason is simply that the interior maximum may result in negative profits, so that the best policy is to produce nothing at all. In other words, production will take place at a positive level if, and only if, the best positive production level results in nonnegative profit.

Let x^* be the output level which is the solution to (5) and satisfies (6). Then x^* will also give a global utility maximum, provided that

$$E[U(px^* - C(x^*) - B)] \geqq U(-B) \tag{22}$$

It will be recalled that $-B$ is the level of profit for $x=0$.[9]

Developing the left-hand side of (22) in a Taylor series around the point $p=\mu$ we obtain, neglecting higher-order terms,

$$E[U(\mu x^* - C(x^*) - B) + U'(\mu x^* - C(x^*) - B)x^*(p-\mu) + \tfrac{1}{2}U''(\mu x^* - C(x^*) - B)x^{*2}(p-\mu)^2] \geqq U(-B)$$

The second term on the left-hand side is zero by definition. Rearranging the remaining terms and dividing through by $U'(\mu x^* - C(x^*) - B)$ so as to make the expressions invariant under linear transformations of the utility function, we then get

$$\frac{U(\mu x^* - C(x^*) - B) - U(-B)}{U'(\mu x^* - C(x^*) - B)} \geqq -\frac{1}{2}\frac{U''(\mu x^* - C(x^*) - B)}{U'(\mu x^* - C(x^*) - B)}x^{*2}E[p-\mu]^2 \tag{23}$$

Both sides of this inequality have the dimension of money. The factors on the right-hand side are the risk aversion function, evaluated at the expected level of profit for $x=x^*$, and the variance of sales, $x^{*2}E[p-\mu]^2$. Since both these factors are positive, the left-hand side must also be positive, and with a strictly increasing utility function this implies that

$$\mu x^* - C(x^*) - B > -B,$$

or

$$\mu > \frac{C(x^*)}{x^*}, \tag{24}$$

i.e., at the optimum *expected price must be larger than average cost, so that the firm requires positive expected profit in order to choose a positive output level.* It should be stressed that "positive" here means "strictly positive." If expected profit for $x=x^*$ were zero, (23) would not be satisfied, and the output level of zero would be chosen. We conclude, therefore, that competitive equilibrium under price uncertainty and risk aversion requires the existence of positive profits.[10]

It is interesting to study the role of risk aversion in the long-run equilibrium posi-

[9] The argument here could equally well be carried out under the "long-run" assumption that $B=0$.

[10] As in any partial equilibrium analysis this statement is somewhat incomplete. Implicit in it is the assumption that by not producing anything the owners of firms can make a sure return by employing their resources elsewhere in the economy. If this return is strictly positive, "normal profits" should be included among the firms' costs.

tion.[11] We assume therefore, to make the discussion simpler, that firms have identical cost functions and identical probability beliefs. Looking at (23) it is easy to see that a ("almost") risk-neutral firm will require only a nonnegative profit to enter the industry; in other words, as long as any positive level of expected profit remains, risk-neutral firms will enter. It is also clear from (23) that firms with "very high" risk aversion will not enter the industry at all, or they will be marginal firms in the sense that a very small decrease in expected price will make them leave the market. The risk neutral firms will of course set marginal cost equal to expected price (assuming U-shaped cost curves), while the risk-averse firms in the industry will choose output levels for which marginal cost is less than expected price. In general, the distribution of output and expected profit among firms will vary with their degree of risk aversion. Expected profit will be highest for those firms which come very close to being risk neutral and have the highest output in the industry. This observation confirms a view which has long traditions in economic theory, viz. to regard profit as a reward to risk-bearing.

Let us now turn to the case where marginal cost is constant or decreasing. We have shown that this case is not inconsistent with competitive assumptions. However, if one or a few firms are much less risk averse than the others, they may choose very high output levels and thereby lower expected price so much that the others will leave the industry. An uneven distribution of risk aversion may therefore be a source of oligopolistic concentration in its own right.

IV. Concluding Remarks

There are many ways in which this analysis can be extended and generalized. We have had nothing to say on the subject of the multiproduct firm, which is of particular interest under uncertainty, since the firm is able to spread its risks by output diversification.[12] Neither have we had anything to say about the role of inventories under demand uncertainty. Finally, investment and financing decisions can hardly be given adequate treatment in the present framework.

It would also be interesting to place the competitive firm facing price uncertainty in a general equilibrium framework. This would require a different type of analysis from that of Debreu, in which there exists a complete set of markets for contingent commodities and the firm bears no risk at all. An alternative approach is contained in a recent paper by Stigum (1969b), in which firms do bear risks and entrepreneurs display risk averse behavior. Evidently, alternative models can be constructed with different assumptions about ownership and market opportunities: the theory of the firm developed in the present paper presumably will fit into some, but not all, of these models.

REFERENCES

K. J. Arrow, *Aspects of the Theory of Risk-Bearing*, Helsinki 1965.

K. H. Borch, *The Economics of Uncertainty*, Princeton 1968.

G. Debreu, *Theory of Value*, New York 1959.

P. J. Dhrymes, "On the Theory of the Monopolistic Multiproduct Firm under Uncertainty," *Int. Econ. Rev.*, Sept. 1964, *5*, 239–57.

J. Drèze and J. J. Gabszewicz, "Demand Fluctuations, Capacity Utilization and Prices," *Operations Research Verfahren*, 1967, *3*, 119–41.

J. Drèze and F. Modigliani, "Consumption Decisions under Uncertainty," CORE Dis-

[11] For the following discussion, which is essentially long-run, it is appropriate to assume $B=0$; in the long run all costs are variable costs.

[12] This problem has been studied by Dhrymes for the special case of a quadratic utility function.

cussion Paper No. 6906, Louvain 1969.

S. H. Hymans, "The Price-Taker: Uncertainty, Utility and the Supply Function," *Int. Econ. Rev.*, Sept. 1966, *7*, 346–56.

H. E. Leland, "The Theory of the Firm Facing Uncertain Demand," mimeo. Stanford Univ., 1969.

———, "On the Existence of Optimal Policies under Uncertainty," mimeo. Stanford Univ., 1970.

J. J. McCall, "Competitive Production for Constant Risk Utility Functions," *Rev. Econ. Stud.*, Oct. 1967, *34*, 417–20.

J. Mossin (1968a), "Aspects of Rational Insurance Purchasing," *J. Polit. Econ.*, July/Aug. 1968, *76*, 553–568.

——— (1968b), "Taxation and Risk-Taking: An Expected Utility Approach," *Economica*, Feb. 1968, *35*, 74–82.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan./Apr. 1964, *32*, 122–136.

A. Sandmo, "The Effect of Uncertainty on Saving Decisions," *Rev. Econ. Stud.*, July 1970, *37*, 353–60.

K. R. Smith, "The Effect of Uncertainty on Monopoly Price, Capital Stock and Utilization of Capital," *J. Econ. Theory*, June 1969, *1*, 48–59.

J. E. Stiglitz, "The Effects of Income, Wealth and Capital Gains Taxation on Risk-Taking," *Quart. J. Econ.*, May 1969, *83*, 263–83.

B. P. Stigum (1969a), "Entrepreneurial Choice over Time under Conditions of Uncertainty," *Int. Econ. Rev.*, Oct. 1969, *10*, 426–42.

——— (1969b), "Competitive Equilibria under Uncertainty," *Quart. J. Econ.*, Nov. 1969, *83*, 533–61.

C. A. Tisdell, *The Theory of Price Uncertainty, Production and Profit*, Princeton 1969.

E. Zabel, "A Dynamic Model of the Competitive Firm," *Int. Econ. Rev.*, June 1967, *8*, 194–208.

The Effect of Tariffs on Production, Consumption, and Trade: A Revised Analysis

By J. CLARK LEITH*

The influence of tariffs on production, consumption, and trade has long occupied a significant place in the literature of international economics. Until recently these influences were analyzed entirely in the context of tariffs on final goods, but the introduction of tariffs on intermediate goods into the discussion has resulted in a renewed interest and revision of the analysis. The new approach has been carried forward in both general and partial equilibrium models, with major interest centering on the partial equilibrium aspects of the stimulus to production in the form of the theory of "effective protection."[1] In this paper, we bring together the production effect of tariffs with the consumption and use effects to focus on the net effect of a set of tariffs on imports of a commodity. The purpose is to show that the assumed elasticity of substitution between inputs in production is a significant element in the partial equilibrium[2] analysis. We specify two alternative models which differ in the elasticity of substitution. In the first, we develop a model under the assumption of a zero elasticity of substitution commonly found in much of the effective protection literature. In the second, we consider the consequences of a model in which there is a unitary elasticity of substitution between inputs in production.

I. Assumptions

Consider the market under free trade for a particular good, J, which is one of many. Assume that foreign produced J is a perfect substitute for the domestically produced J, and that the foreign supply is infinitely elastic at a price that is below the autarky price but not below the zero offer price of domestic production. Under free trade, domestic producers offer that amount at which the world price equals their marginal costs, consumers and users purchase that amount at which the world price equals the marginal usefulness of J to them, and the difference is made up of imports.

If we now impose a set of tariffs affecting the market for good J, the markets for J's inputs, and the markets where J is used as an input, then domestic production

* Assistant professor, department of economics, University of Western Ontario, and currently visiting at the University of Ghana under a twinning arrangement between the two departments. He is indebted to W. M. Corden, Herbert Grubel, Peter J. Lloyd, James R. Melvin, P. O'Brien, J. R. Williams, and G. D. Wood for helpful comments on an early draft. None, however, should be held responsible for what is presented here.

[1] The seminal partial equilibrium contributions are by W. M. Corden (1966) and Harry Johnson (1965). Corden also makes a beginning in the task of extending the partial equilibrium analysis in the direction of a general equilibrium system. And, there are the important early contributions to general equilibrium analysis of Ronald McKinnon and William Travis (1964). More recent general equilibrium contributions include Roy Ruffin, and V. K. Ramaswami and T. N. Srinivasen.

[2] The analysis is restricted to partial equilibrium for a number of general equilibrium effects are ignored, including cross-elasticities of demand effects, income effects, the effects of changes in factor prices on other industries, and the effects of tariffs on the balance of trade and the exchange rate.

consumption, use, and hence imports will be changed. In analyzing these changes, assume further that there is domestic production and trade both with and without tariffs, and that there is a production function homogeneous of degree one for each good that relates inputs of tradeable materials and a primary factor to output.[3]

II. Zero Elasticity of Substitution Case[4]

First, consider the effect of tariffs on production of J when we assume a zero elasticity of substitution between all inputs used to produce it. The change in production of J is[5]

$$dS_j = \left(t_j - \sum_i a_{ij}t_i\right)e_jS_j \tag{1}$$

where I is the typical input, and the a_{ij} coefficient is the value of I used per dollar of J output, at free trade prices. The term $\sum_i a_{ij}t_i$ describes the upward shift in the supply curve due to the tariffs on inputs, and the difference between that and the price increase of the output $(t_j - \sum_i a_{ij}t_i)$ is the *net* rate of protection received by producers of good J.[6] It is the net protection, together with the elasticity of supply of the original supply curve e_j, that describes the expansion of domestic production. Equation (1) can be simplified by replacing $(t_j - \sum_i a_{ij}t_i)$ with the single symbol π_j representing the net rate of protection afforded domestic output by the tariff structure:

$$dS_j = \pi_je_jS_j \tag{2}$$

The change in household consumption is

$$dC_j = \eta_jt_jC_j \tag{3}$$

and the change in intermediate use is

$$\begin{aligned} dU_j &= \sum_i a_{ji}\left(t_i - \sum_j a_{ji}t_j\right)e_iS_i \\ &= \sum_i a_{ji}\pi_ie_iS_i \end{aligned} \tag{4}$$

which describes the shift of the demand curve, for $\pi_ie_iS_i$ is the expanded production in the typical using industry, I, and the a_{ji} term indicates the impact this has on industry J.

The change in imports is the sum of (3) and (4) minus (2).

$$dM_j = -\pi_je_jS_j + \eta_jt_jC_j + \sum_i a_{ji}\pi_ie_iS_i, \tag{5}$$

or, to be explained below

$$dM_j = -f_j\epsilon_jS_j + \eta_jt_jC_j + \sum_i a_{ji}f_i\epsilon_iS_i$$

Our attention so far has been confined to influence of tariffs on the market for the typical *product*. We can now indicate how this analysis corresponds to the idea embodied in the effective protection concept of protection of a *process*. The effective rate of protection is the proportionate change in the price of the primary factor that is made possible by the imposition of tariffs (f_j). The formula is

$$f_j = \frac{t_j - \sum_i a_{ij}t_i}{1 - \sum_i a_{ij}} = \frac{\pi_j}{v_j}, \tag{6}$$

[3] The assumption of a single primary factor is clearly an oversimplification. However, without being excessively arbitrary, it can be taken to represent a composite of primary factors, and changes in its price as the weighted average change in the price of the primary factors. See Lloyd for an approach that uses multiple primary inputs.

[4] This section draws on the analysis developed by Rachel Dardis, and Johnson (1969). The latter, in turn, has been used by Bela Balassa (1965, 1967), and Leith and G. L. Reuber in discussing the restriction of imports due to tariffs. While it could be argued that the shifts of the demand and supply curves described here are not partial equilibrium effects in the sense that the influence of more than a change in the price of J is considered, we do confine our attention to the effects on this *one* market of *one set* of tariffs. We ignore effects such as changes in relative prices of other substitutes and complements on both the production and consumption sides.

[5] Throughout the paper we will utilize linear approximations for discrete changes along supply and demand curves.

[6] The concept of the net protection received by producers was used by Paul and Ronald J. Wonnacott.

where v_j is value-added per dollar of output of industry J. This is clearly related to the production effect indicated in equations (1) and (2) above, and it can be shown that the effective rate of protection together with the elasticity of supply of the primary factor is an equivalent way of representing the production effect. Let S_{fj} represent the quantity of the primary factor used in production of J, ϵ_j the elasticity of supply of the primary factor used in the production of J, and α_{fj} is the quantity of the primary factor that is used per unit of output in the free trade equilibrium situation. Thus, $\alpha_{fj}=S_{fj}/S_j$, and under free trade

$$(7) \qquad S_j = S_{fj}/\alpha_{fj},$$

and with protection[7]

$$(8) \quad S_j(1 + dS_j/S_j) = \frac{S_{fj}(1 + dS_{fj}/S_{fj})}{\alpha_{fj}(1 + d\alpha_{fj}/\alpha_{fj})}$$

Because of zero substitution, $d\alpha_{fj}/\alpha_{fj}=0$. Now define

$$\epsilon_j \equiv \frac{dS_{fj}/S_{fj}}{f_j}$$

Therefore (8) may be written

$$(9) \qquad S_j(1 + dS_j/S_j) = \frac{S_{fj}}{\alpha_{fj}}(1 + \epsilon_j f_j),$$

but we may eliminate the initial conditions (7) to obtain

$$(9') \qquad dS_j = \epsilon_j f_j S_j$$

Equation (9′) is thus equivalent to equation (2) as a way of representing the expansion of domestic output.

Although either the approach of equation (9′) or that of equation (2) tells us by how much output changes, we do not have this choice of approaches when we consider the influence of tariffs on consumption and imports. Consumption and trade can be measured only in units of the product, not units of the process.

Consider also the pull of resources into the production of good J due to protection.[8] The proportionate change in use by industry J of the typical input I (dS_{ij}/S_{ij}) is readily specified under the fixed coefficient assumption: the proportionate change in use of any input is equal to the proportionate change in output. Hence

$$(10) \qquad \frac{dS_{ij}}{S_{ij}} = \frac{dS_j}{S_j} = \pi_j e_j,$$

where I is any input, material or primary. Again, the net rate of protection and the supply elasticity must be known. From this we can determine the proportionate change in the price of the typical input I (dp_{ij}/p_{ij}) induced by the expansion of output

$$(11) \qquad dp_{ij}/p_{ij} = \pi_j(e_j/e_{ij})$$

where e_{ij} is the elasticity of supply of input I to industry J. Under the assumption of infinitely elastic supplies of tradeables, the proportionate change in the price of a tradeable input is clearly zero. However, if I is a non-traded material input, the change in its price is given by equation (11).[9] The proportionate change in the price of the primary factor (i.e., the effective rate of protection), is a special case of (11):

$$(11') \qquad dp_{ij}/p_{ij} = \pi_j(e_j/\epsilon_j) \quad \text{(for } I = \text{primary factor input)},$$

and since $\epsilon_j=v_j e_j$

$$(11'') \qquad dp_{ij}/p_{ij} = \pi_j/v_j \quad \text{(for } I = \text{primary factor input)}.$$

[7] Note that whether or not the quantity of the primary factor used per unit of output will change depends on whether there is substitution of inputs.

[8] The concepts that follow here draw in part on Benton Massell.

[9] We must assume, however, that there are no tariffs on inputs used in the production of non-traded inputs. (See Leith (1968), p. 591.)

Finally, consider the effect of tariffs on domestic costs under protection. The stimulus to production is π_j (or f_j), but the higher marginal cost of domestic production is still given by the nominal tariff.[10] This is made up of the higher cost of inputs due to tariffs on them, plus the induced higher prices of primary and non-traded material inputs, the induced portion being shared in proportion to the inputs' supply elasticities. This may be seen by rearranging (11) and summing over all inputs, yielding

$$t_j = \sum_i \left(a_{ij} t_i + \frac{dp_{ij}}{p_{ij}} \cdot \frac{e_{ij}}{e_j} \right) \qquad (12)$$

(for I = all inputs).

III. Unitary Elasticity of Substitution Case

Consider now the effect of tariffs on production, consumption, use, and trade when we assume, instead of a zero elasticity of substitution, a unitary elasticity of substitution between inputs in production.[11]

The change in production that is due to tariffs is most easily seen by relating the change in the price of the primary factor to the expansion of domestic production. Recall that the free trade prices of materials and the output are pegged by the world market prices, and that it is the elasticity of supply of the primary factor that constrains the expansion of domestic production. In the zero substitution case one way of specifying the expansion of domestic production was by the following (notional) three step procedure: a) determine the effective rate of protection, i.e., the proportionate change in the price of the primary factor (equation (6)); b) determine the proportionate change in the quantity supplied of the primary factor from the elasticity of supply of the primary factor and the effective rate of protection; and c) determine the proportionate change in the quantity of output (equation (9)), given the proportionate change in the quantity supplied of the primary factor and the proportionate change in the number of units of the primary factor per unit of output, the latter being zero in the case of no substitution.

Consider a similar three step procedure in the case of production function homogeneous of degree one with unitary elasticity of substitution, i.e., a Cobb-Douglas production function. First, to calculate the proportionate change in the price of the primary factor, take the Cobb-Douglas total cost function

$$S_j p_j = \frac{\left(\prod_i p_i^{a_{ij}} \right) p_f^{v_j} S_j}{k} \qquad (13)$$

where $k = K(a_{ij}^{a_{ij}} \cdot v_j^{v_j})$, which is a constant, and p represents per unit price. Rearranging (13)

$$p_j k = \left(\prod_i p_i^{a_{ij}} \right) p_f^{v_j}, \qquad (14)$$

and since our concern is with relative price changes, set $p_j k = 1$, $p_i = 1$ for all i, and $p_f = 1$. Inflating the various prices due to tariffs

$$1 + t_j = \left[\prod_i (1 + t_i)^{a_{ij}} \right] (1 + f_j)^{v_j} \qquad (15)$$

[10] This holds regardless of the degree of substitution in production. Note also that discovery of input tariffs and the effective rate of protection does not mean discovery of any heretofore hidden costs due to tariffs. Rather, we are better able to disentangle the distribution of those cost increases permitted by the nominal tariff. Contrast, for example, Arthur Smith in the introduction of J. R. Melvin and B. W. Wilkinson: "The effective tariff rates derived in this study suggest that cost and productivity effects of the tariff structure in Canada may be significantly larger than those indicated by the nominal rates."

[11] The question of substitution was raised initially by Corden (1966) and Travis (1964). Subsequent contributions included: James Anderson and Seija Naya, Corden forthcoming, Grubel and Lloyd, Leith (1968), Lloyd, Ramaswami and Srinivasen, Augustine H. H. Tan, and Travis (1968).

and solve for f_j[12]

$$f_j = \left[\frac{1+t_j}{\prod_i (1+t_i)^{a_{ij}}}\right]^{1/v_j} - 1 \tag{16}$$

Second, given f_j from (16), and ϵ_j, dS_{fj}/S_{fj} is obtained. Third, determine $d\alpha_{fj}/\alpha_{fj}$, and substitute both dS_{fj}/S_{fj} and $d\alpha_{fj}/\alpha_{fj}$ into (9) to obtain dS_j/S_j. The expression for $d\alpha_{fj}/\alpha_{fj}$ is obtained by expanding the relationship

$$\alpha_{fj} = v_j \frac{p_j}{p_f} \tag{17}$$

With the introduction of tariffs, (17) becomes

$$\alpha_{fj}(1 + d\alpha_{fj}/\alpha_{fj}) = \frac{v_j p_j (1+t_j)}{p_f(1+f_j)} \tag{18}$$

Since (17) holds, cancelling on both sides of (18) yields

$$1 + d\alpha_{fj}/\alpha_{fj} = \frac{(1+t_j)}{(1+f_j)} \tag{19}$$

And hence from (19)

$$d\alpha_{fj}/\alpha_{fj} = \frac{1+t_j}{1+f_j} - 1 \tag{20}$$

Thus, for the Cobb-Douglas case[13] equation (9) becomes

$$\begin{aligned} dS_j/S_j &= \frac{1+\epsilon_j f_j}{(1+t_j)/(1+f_j)} - 1 \\ &= \frac{(1+\epsilon_j f_j)(1+f_j)}{1+t_j} - 1 \end{aligned} \tag{21}$$

When substitution is allowed, economizing of more expensive inputs becomes possible. To illustrate the quantitative significance of substitution in determining the magnitude of the expansion of domestic production, a numerical example is contained in Table 1. Economizing takes place in the unitary elasticity of substitution case ($\sigma = 1$) whenever *relative* prices of inputs change. Thus, in every case, except where all tariffs are the same leaving relative prices of all inputs unchanged (col. 5), the effective rate of protection and the change in production are different between $\sigma = 1$ and $\sigma = 0$. Further, note that when the tariff rate on the output (t_j) exceeds the average tariff rate on the inputs ($\bar{t}_i$, where $\bar{t}_i = \sum_i a_{ij}t_i/a_{ij}$), there is economizing of the primary factor, and where $t_j < \bar{t}_i$, relatively more of the primary factor is used.[14] As a result of the economizing, the absolute rate of change in domestic output is greater where economizing takes place than when it is not permitted. Note also that the combination of tariffs that yields a zero effective rate of protection is different between the two cases: for $\sigma = 0$, $f_j = 0$ when $t_j = \sum a_{ij}t_i$ (col. 4); and for $\sigma = 1$, $f_j = 0$ when

$$(1 + tj) = \prod_i [(1+t_i)^{a_{ij}}]$$

[12] Note that allowing a positive elasticity of substitution ($\sigma > 0$) between inputs results in a higher effective rate of protection than in the zero elasticity of substitution case, where the effective rate of protection is defined as the proportionate change in the price of the primary factor input. (See Leith (1968) p. 600.) If, however, the effective rate of protection is defined as the proportionate change in per unit value-added, no statement can be made about the bias introduced by allowing substitution unless the tariffs and elasticity of substitution are known. And, if the effective rate of protection is defined in this way, it is no longer an indicator of resource pulls. This is a further reason for confining our attention to partial equilibrium analysis. In a general equilibrium model neither the price nor the per unit value-added definition of the effective rate of protection is an indicator of resource pull. The value-added definition in a general equilibrium model suffers from the same problem arising from substitution as in the partial equilibrium model. The price definition has no meaning in a general equilibrium context because domestic factor prices are identical between all activities with and without protection, and hence the proportionate changes in the factor prices are identical between all activites.

[13] The same approach can be followed using the effective rate of protection calculated from the *CES* production function. (See Leith 1968, Part iii.) The procedure is: (a) calculate the effective rate of protection using equation (12) in ibid.; and (b) determine da_{ij}/a_{ij} working from the input demand equation (10) in ibid.; and (c) calculate dS_j/S_j from equation (9) of this paper.

[14] For a geometric representation of substitution between the primary factor and material inputs, see Leith (1967.)

TABLE 1—NUMERICAL EXAMPLE OF ERP AND PRODUCTION CHANGE UNDER ZERO AND UNITARY ELASTICITY OF SUBSTITUTION ASSUMPTIONS

(Free trade coefficients: $a_{1j}=.33$, $a_{2j}=.33$, $v_j=.33$, and assumed $\epsilon_j=.33$)

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
t_j	output	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20
t_1	input 1	.50	.40	.50	.30	.20	.10	.10	.05	0	−.10
t_2	input 2	.50	.40	.30	.30	.20	.10	0	.05	0	+.10
$\bar{t}_i$	av. input tariff	.50	.40	.40	.30	.20	.10	.05	.05	0	0
f_j	$\sigma=0$	−.400	−.20	−.20	0	.20	.40	.50	.50	.60	.60
f_j	$\sigma=1$	−.232	−.118	−.114	+.0225	.20	.428	.571	.567	.728	.745
$d\alpha_{ij}/\alpha_{ij}$	$\sigma=0$	0	0	0	0	0	0	0	0	0	0
$d\alpha_{ij}/\alpha_{ij}$	$\sigma=1$	+.563	+.361	+.354	+.174	0	−.160	−.236	−.234	−.306	−.312
dS_j/S_j	$\sigma=0$	−.133	−.067	−.067	0	+.067	+.133	+.167	+.167	+.20	+.20
dS_j/S_j	$\sigma=1$	−.410	−.294	−.290	−.141	+.067	+.361	+.558	+.552	+.791	+.814

Note: $\bar{t}_i$ is the weighted average input tariff: i.e., $\bar{t}_i = \sum_i a_{ij}t_i \Big/ \sum_i a_{ij}$

Where there are input tariffs two effects must be distinguished: substitution between material and primary inputs; and substitution between material inputs. If all material inputs are subject to the same tariff, the substitution is only between the factor and the materials, and not between materials (e.g., col. 8). However, where inputs are subject to different tariffs, there is economizing of the relatively more expensive inputs, and the cost-increasing effect of input tariffs is not as strong, the effective rate of protection is (algebraically) greater, and the change in production is (algebraically) greater (e.g., col. 7 vs. 8, and col. 3 vs. 2). Further, an import subsidy on one input to offset an import tariff on another is *not* equivalent to identical tariffs on both inputs yielding the same average input tariffs. For example, $t_1=t_2=0$ (col. 9) is not equivalent to $t_1=-.10$ and $t_2=.10$ so that $\bar{t}_i=0$ (col. 10), for substitution in the latter case between inputs 1 and 2 means that the effective rate of protection is greater and the expansion of production is greater than in the former case.

In the expansion of J production we draw not only primary factors, but also material inputs. In a manner similar to the derivation of (21) we can solve for $d\alpha_{ij}/\alpha_{ij}$ from the relationship $\alpha_{ij}=a_{ij}p_j/p_i$ with the result

$$\frac{d\alpha_{ij}}{\alpha_{ij}} = \frac{1+f_j}{1+t_i} - 1 \tag{22}$$

Then, from $S_{ij}=S_j\alpha_{ij}$ we can solve for

$$\frac{dS_{ij}}{S_{ij}} = \frac{(1+\epsilon_j f_j)(1+f_j)}{1+t_i} - 1 \tag{23}$$

In addition to the change in production, consider now the change in consumption and the change in use of good J in the case of unitary rather than zero elasticity of substitution.

Clearly the change in consumption is unaffected by the elasticity of substitution in production. Thus, equation (3) still holds.

The change in use equation, however, is altered by the introduction of substitution because the drawing of J into the various other I activities is affected by the elas-

ticity of substitution in the latter's production functions. Thus, in a manner similar to (23), and with some rearranging, the change in use of good J is

$$(24)\quad dU_j = \sum_i a_{ji}\left[\frac{(1+\epsilon_i f_i)(1+f_i)}{1+t_j} - 1\right]S_i$$

Finally, summing all the components of the change in imports, when $\sigma=1$,

$$dM_j = -\left[\frac{(1+\epsilon_j f_j)(1+f_j)}{1+t_j} - 1\right]S_i$$

$$(25)\quad + \eta_j t_j C_j$$

$$+ \sum_i a_{ji}\left[\frac{(1+\epsilon_i f_i)(1+f_i)}{1+t_j} - 1\right]S_i$$

Comparing the change in imports due to tariffs in the case of $\sigma=1$ with the case of $\sigma=0$ shows that: a) the component measuring the change in domestic output of J will be greater for $\sigma=1$ when $f_j > t_j$; b) the component measuring the change in consumption will be the same between $\sigma=1$ and $\sigma=0$; and c) the component measuring the change in use of J by the typical using industry I will be greater for $\sigma=1$ when $f_i > t_j$.

REFERENCES

J. Anderson and S. Naya, "Substitution and Two Concepts of Effective Rate of Protection," *Amer. Econ. Rev.*, Sept. 1969, *59*, 607–12.

B. Balassa, "The Impact of the Industrial Countries' Tariff Structure on Their Imports of Manufactures from Less Developed Areas," *Economica*, Nov. 1967, *34*, 372–83.

———, "Tariff Protection in Industrial Countries: An Evaluation," *J. Polit. Econ.*, Dec. 1965, *73*, 572–94.

W. M. Corden, "The Structure of a Tariff System and the Effective Protective Rate," *J. Polit. Econ.*, June 1966, *74*, 221–37.

———, "The Substitution Problem in the Theory of Effective Protection," in *Theory of Protection*, forthcoming.

R. Dardis, "Intermediate Goods and the Gain from Trade," *Rev. Econ. Statist.*, Nov. 1967, *49*, 502–09.

H. G. Grubel and P. J. Lloyd, "Factor Substitution and Effective Tariff Rates," *Rev. Econ. Stud.*, forthcoming.

H. G. Johnson, "The Theory of Tariff Structure with Special Reference to World Trade and Development," in *Trade and Development*, Geneva 1965.

———, "The Theory of Effective Protection and Preferences," *Economica*, May 1969, *36*, 119–38.

J. C. Leith, "Substitution and Supply Elasticities in Calculating the Effective Protective Rate," *Quart. J. Econ.*, Nov. 1968, *82*, 588–601.

———, "Effective Rates of Protection: Analysis and an Empirical Test," unpublished doctoral dissertation, Univ. Wis., 1967.

——— and G. L. Reuber, "The Impact of the Industrial Countries' Tariff Structure on their Imports of Manufacturers from Less Developed Areas: A Comment," *Economica*, Feb. 1969, *36*, 75–80.

P. J. Lloyd, "Effective Protection: A Different View," *Econ. Rec.*, forthcoming.

R. I. McKinnon, "Intermediate Products, and Differential Tariffs: A Generalization of Lerner's Symmetry Theorem," *Quart. J. Econ.*, Nov. 1966, *80*, 584–615.

B. F. Massell, "The Resource-Allocative Effects of a Tariff and the Effective Protection of Individual Inputs," *Econ. Rec.*, Sept. 1968, *44*, 369–76.

J. R. Melvin and B. W. Wilkinson, *Effective Protection in the Canadian Economy*, Economic Council of Canada Special Study No. 9, Ottawa 1968.

V. K. Ramaswami and T. N. Srinivasan, "Tariff Structure and Resource Allocation in the Presence of Factor Substitution," in J. N. Bhagwati et al, eds., *Trade, Balance of Payments, and Growth; Essays in Honor of Charles Kindleberger*, Cambridge, Mass. 1971.

R. J. Ruffin, "Tariffs, Intermediate Goods, and Domestic Protection," *Amer. Econ. Rev.*, June 1969, *59*, 261–69.

A. H. H. Tan, "Differential Tariffs, Negative Value-Added and the Theory of Effective

Protection," *Amer. Econ. Rev.*, Mar. 1970, *60*, 107–16.

W. P. Travis, *The Theory of Trade and Protection*, Cambridge, Mass. 1964.

———, "The Effective Rate of Protection and the Question of Labor Protection in the United States," *J. Polit. Econ.*, May-June 1968, *76*, 443–61.

P. Wonnacott and R. J. Wonnacott, *Free Trade Between the United States and Canada: The Potential Economic Effects*, Cambridge, Mass. 1967.

A General Disequilibrium Model of Income and Employment

By Robert J. Barro and Herschel I. Grossman*

As is now well understood, the key to the Keynesian theory of income determination is the assumption that the vector of prices, wages, and interest rates does not move instantaneously from one full employment equilibrium position to another. By implication, Keynesian economics rejects the market equilibrium framework for analyzing the determination of quantities bought, sold, and produced. This framework is associated with Walras and Marshall, both of whom proceeded as if all markets were continuously cleared. Walras rationalized this procedure by incorporating recontracting arrangements, while Marshall did so by regarding price adjustments to be an instantaneous response to momentary discrepancies between quantities supplied and demanded.

By rejecting these rationalizations, Keynesian theory proposes as a general case a system of markets which are not always cleared. Keynes was, tacitly at least, concerned with the general theoretical problem of the intermarket relationships in such a system. The failure of a market to clear implies that, for at least some individuals, actual quantities transacted diverge from the quantities which they supply or demand. Thus, the natural focus of Keynesian analysis is on the implications for behavior in one market of the existence of such a divergence in another market. Indeed, some recent writers, such as Robert Clower and Axel Leijonhufvud, have argued very convincingly that this focus is the crucial distinguishing feature of Keynesian economics.

Unfortunately, the evolution of conventional post-Keynesian macroeconomics failed to interpret the Keynesian system in this light.[1] Instead, conventional analysis has chronically attempted to coax Keynesian results out of a framework of general market equilibrium. The result has been to leave conventional macroeconomics with an embarrassingly weak choice-theoretic basis, and to associate with it important implications which are difficult to reconcile with observed phenomena.

A classic example of such a difficulty concerns the relationship between the level of employment and the real wage rate. In the conventional analysis, the demand for labor is inversely and uniquely related to the level of real wages. This assumption accords with Keynes; who, in this respect, had adhered to received pre-Keynesian doctrine.[2] Given this assumption, cyclical variations in the quantity of labor demanded and the amount of employment must imply countercyclical variation in real wage rates. As is well known, however, such a pattern of real wages has not been observed.[3]

* Assistant professor and associate professor of economics, respectively, Brown University. National Science Foundation Grants GS-2419 and GS-3246 supported this research.

[1] See Leijonhufvud.

[2] Keynes wrote:

> ... with a given organization, equipment and technique, real wages and the volume of output (and hence of employment) are uniquely correlated, so that, in general, an increase in employment can only occur to the accompaniment of a decline in the rate of real wages. Thus, I am not disputing this vital fact which the classical economists have (rightly) asserted.... The real wage earned by a unit of labor has a unique inverse correlation with the volume of employment. [1936, p. 17]

[3] The evidence has been recently reviewed by Edwin

A few authors have pointed out the inappropriateness of attempts to force Keynesian analysis into a market equilibrium framework. Contributions by Don Patinkin (1956) and Clower, in particular, represent important attempts to reconstruct macroeconomic theory within an explicitly disequilibrium context.

In the unfortunately neglected chapter 13 of *Money, Interest, and Prices*, Patinkin analyzed involuntary unemployment in a context of explicit market disequilibrium; and he showed that the misleading implications of the conventional analysis regarding the real wage are a direct consequence of its general equilibrium character.[4] Patinkin presented a theory in which involuntary unemployment of labor can arise as a consequence of disequilibrium, in particular, excess supply in the market for current output. In this theory, the inability of firms to sell the quantity of output given by their supply schedule causes them to demand a smaller quantity of labor than that given by their conventional (or notional) demand schedule. The immediate significance of this theory is that it is able to generate unemployment without placing any restrictions on the level or movement of the real wage.[5] Unemployment of labor requires only that the vector of prices and wages implies a deficiency of demand for current output. As Patinkin suggests, this interpretation of the proximate cause of unemployment is more Keynesian than Keynes' own discussion.

The essence of Patinkin's theory is causality running from the level of excess supply in the market for current output to the state of excess supply in the market for labor. Patinkin thereby explains the proximate cause of cyclical unemployment, but his analysis involves only partial, rather than general, disequilibrium. At the least, a general disequilibrium model would, in addition, incorporate the possibility of a reverse influence of the level of excess supply in the labor market upon the state of excess supply in the market for current output.

Clower's important paper develops a theory emphasizing this causal relationship. He presents a derivation of the Keynesian consumption function in which he interprets the relationship between consumption and income as a manifestation of disequilibrium in the labor market. This approach to explaining household behavior is obviously similar to Patinkin's analysis of the firm. The only significant difference is that Clower's households have a choice between consuming and saving, so that his problem is explicitly choice theoretic. However, if Patinkin's approach were generalized to a multi-input production function, the resulting analysis would be formally analogous to Clower's.

The analysis in this paper builds on the foundations laid down by the Patinkin and Clower analyses of a depressed economy. Our purpose is to develop a generalized analysis of both booms and depressions as disequilibrium phenomena.[6] Section I

Kuh, esp. pp. 246–48; and Ronald Bodkin. Keynes (1939) recognized this discrepancy, and offered a rather contrived explanation for it in terms of monopoly and procyclical variation in demand elasticities. More recently, Kuh attempted to explain this discrepancy in terms of a fixed proportions production function in the short run.

[4] Chapter 13 also appears, apparently unchanged, in the second edition of *Money, Interest, and Prices* (1965). Patinkin had first presented some of the essentials of this analysis in an earlier article (1949). A similar formulation appears in Edgar Edwards.

[5] Patinkin's theory does not involve any restrictions either upon the substitutability among factors of production or upon demand elasticities. (See fn. 3.) Of course, this theory does not deny that an excessive level of real wages can be an independent cause of unemployment. But, a clear analytical distinction is made between unemployment due to this cause, and unemployment which occurs even when the level of real wages is not excessive.

[6] The analysis by Robert Solow and Joseph Stiglitz, although they emphasize different questions, is somewhat similar to the present approach. However, their analytical format does differ from ours in at least three

sketches the analytical framework employed. Section II reviews and generalizes Patinkin's analysis of the labor market and involuntary unemployment. Section III develops a distinction, implied by Patinkin's analysis, between two concepts of unemployment; one associated with excess supply in the labor market and the other associated with equilibrium in the labor market but with disequilibrium elsewhere in the system. Section IV reviews Clower's analysis and shows how it is formally analogous to Patinkin's. Section V joins the Patinkin and Clower analyses into a model of an economy experiencing deficient aggregate demand. Section VI formulates an analogous model of an economy experiencing excessive aggregate demand. Finally, Section VII summarizes the main results.

I. Analytical Framework

The following discussion utilizes a simple aggregative framework which involves three economic goods—labor services, consumable commodities, and fiat money—and two forms of economic decision making unit—firms and households. Labor services are the only variable input into the production process. Other inputs have a fixed quantity, no alternative use, and zero user cost. Consumable commodities are the only form of current output; there is no investment.[7] Money is the only store of value, and it also serves as a medium of exchange and unit of account. The nominal quantity of money is exogenous and constant.

Firms demand labor and supply commodities. They attempt to maximize profits. Households supply labor and demand commodities and money balances. They also receive the profits of the firms according to a predetermined distribution pattern. Households attempt to maximize utility. Each firm and household is an atomistic competitor in the markets for both commodities and labor.

Following Patinkin (1956, 1965), each of the flow variables in the model—commodities, labor services, and the increment to money balances—is for simplicity expressed as the quantity which accrues over a finite unit of time, say a week, so that each assumes the dimensions of a stock. The model thus includes the following variables:

y = quantity of commodities
x = quantity of labor services
m = increment to real money balances (in commodity units)
π = quantity of real profits (in commodity units)
M = initial stock of nominal money balances
P = money price of commodities
w = real wage rate (in commodity units)

Throughout the following discussion, the method of analysis is to take a particular vector of the price level and real wage rate as given, and to work out the levels of income and employment implied by that vector. This procedure represents a non-Marshallian, or Keynesian, extreme, and following John Hicks may be denoted as the "fix-price method." The analysis does, of course, have implications for the appropriate specification of the forces making for changes in prices and wages. This paper does not explicitly investigate these implications, although we do consider a parenthetical example concerning the model's implications for the cyclical behavior of real wages.[8]

substantial respects: First, they do not discuss the choice-theoretic basis for the theory. Second, the equilibrium price level is indeterminate in their model. Third, by introducing restrictions on the rate of change of employment, they complicate matters and obscure what would seem to be essential in the intermarket effects of disequilibrium.

[7] It should be clear that the incorporation of investment and a market for securities would alter none of the conclusions advanced in this paper.

[8] Grossman develops a more general model of multimarket disequilibrium based on Clower's choice-theoretic paradigm, and focuses in detail on the implications

II. Patinkin's Analysis of the Labor Market

Consider the behavior of the representative firm under the provisional assumption that it regards profit maximization as being constrained only by the production function. In particular, the firm perceives that it can purchase all the labor which it demands and sell all the output which it supplies at the existing levels of w and P. Thus, profits are given by

$$\pi = y^S - wx^D,$$

where the superscripts indicate supply and demand quantities. Assuming the production function to be

$$y = F(x),$$

with positive and diminishing marginal product, profit maximization implies

$$x^D = x^D(w),$$

such that $\partial F/\partial x = w$, and

$$y^S = F(x^D)$$

Patinkin (1956, 1965) contrasts the above to a situation in which commodities are in excess supply. Voluntary exchange implies that actual total sales will equal the total quantity demanded. The representative firm will not be able to sell its notional supply y^S.[9] Let y represent its actual demand-determined sales, where $y < y^S$.[10] Then, the profit maximization problem becomes simply to select the minimum quantity of labor necessary to produce output quantity y.[11] In other words, the firm maximizes

$$\pi = y - wx^{D'},$$

subject to $y = F(x)$. The variable $x^{D'}$ may be denoted as the effective demand for labor. Profit maximization now implies

$$(1) \qquad x^{D'} = F^{-1}(y) \quad \text{for } \frac{dF}{dx} \geq w$$

The constraint of $y < y^S$ implies $x^{D'} < x^D$, with $x^{D'}$ approaching x^D as y approaches y^S.[12]

of this model for the disequilibrium behavior of prices and interest.

[9] We assume here that the firm would actually like to sell y^S. Such behavior may not always be optimal. For example, Section VI discusses a situation of excess demand for labor in which the firm's effective supply $y^{S'}$ is less than y^S. However, we assume for simplicity that excess demand for labor never coexists with excess supply of commodities and vice versa. Grossman presents a more general treatment of multi-market disequilibria which allows for the coexistence of excess supply in one market and excess demand in another, as well as excess supply or demand in both.

[10] In principle, y need not be less than y^S for every firm. The apportionment of the actual sales among the firms depends upon established queuing or rationing procedures. Grossman presents an explicit analysis of this apportionment within a framework of voluntary exchange.

The inability of a firm to sell its desired output at the going price violates an assumption of the perfectly competitive model. Kenneth Arrow has stressed this inconsistency of perfect competition with disequilibrium. Essentially, he argues that economic units which act as perfect competitors in equilibrium must (at least in certain respects) perform as monopolists in disequilibrium. In this paper we focus on the reaction of economic units to given (equilibrium or disequilibrium) price levels. If, in addition, one wished to analyze explicitly the dynamics of price adjustment, it would be necessary to discard the perfectly competitive paradigm of the producer as a price taker. (In this regard, see Barro 1970, 1971.)

[11] This analysis abstracts from inventory accumulation or decumulation. For simplicity, we assume throughout that output always adjusts instantaneously to equal the smaller of supply and demand. Permitting inventory accumulation would not affect the essentials of the analysis, although it would introduce a complication analogous to the inclusion of an additional input. In general, we might obtain $dy/dt = k[\min(y^D, y^S) - y]$, where $k = k(w, y) > 0$. A similar gradual adjustment process for employment might also be possible, as in Solow and Stiglitz.

[12] The choice-theoretic nature of the problem becomes much more interesting when there is more than one form of input. Assume profits to be given by $\pi = y - w_1 x_1^{D'} - w_2 x_2^{D'}$, where the production function is $y = F(x_1, x_2)$, which has the usual convexity properties. Profit maximization now implies

$$(1.1) \qquad x_1^{D'} = x_1^{D'}\left(\frac{w_1}{w_2}, y\right)$$

$$(1.2) \qquad x_2^{D'} = x_2^{D'}\left(\frac{w_1}{w_2}, y\right)$$

such that at output y, $(\partial F/\partial X_1)/(\partial F/\partial X_2) = (w_1/w_2)$. In reducing output y^S to y, the firm must now make a decision regarding optimal input combinations. However, as y approaches y^S, $x_1^{D'}$ and $x_2^{D'}$ approach x_1^D and x_2^D.

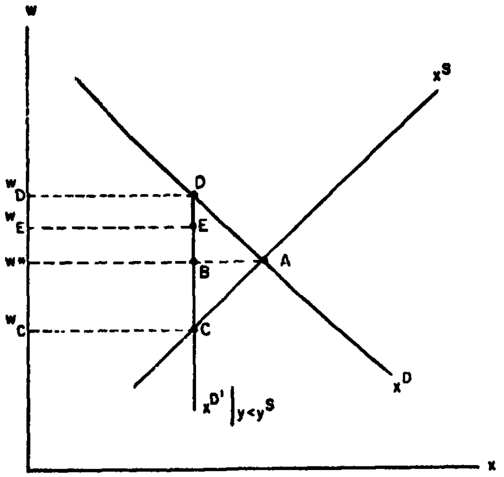


FIGURE 1. THE LABOR MARKET WITH EXCESS SUPPLY OF COMMODITIES

The essential implication of equation (1) is that the effective demand for labor can vary even with the real wage fixed. Given voluntary exchange, employment cannot exceed the effective demand for labor. The quantity of employment thus is not uniquely associated with the real wage.

III. The Concept of Unemployment

Figure 1 depicts the preceding analysis of the labor market. The notional demand schedule for labor x^D is downward sloping. If $y=y^S$, the effective demand for labor $x^{D'}$ coincides with the notional demand. If $y<y^S$, the effective demand is independent of the real wage and less than the notional demand. The (notional) supply schedule for labor x^S, which will be derived below, is shown as upward sloping.

Figure 1 suggests a distinction between two concepts of unemployment—involuntary unemployment associated with excess (effective) labor supply, and voluntary unemployment associated with equilibrium in the labor market, but with disequilibrium elsewhere in the system. Suppose that initially the commodity market is in equilibrium, so that $y=y^S$ and $x^{D'}=x^D$, and that initially the real wage is w^*. Thus, the labor market is in equilibrium at point A, which may be denoted as full employment general equilibrium. Now suppose, say because the price level P is too high, that commodity demand is lower so that $y<y^S$ and $x^{D'}<x^D$. At the real wage w^*, excess supply of labor will amount to quantity AB. Failure of the price level to adjust to clear the commodity market leads to excess supply in the labor market. This excess supply represents what we usually refer to as involuntary unemployment. It is also what the Bureau of Labor Statistics ideally intends to represent by its statistical measure of unemployment—those seeking but not obtaining work at the going real wage. Involuntary unemployment clearly does not require a rise in the real wage above the level consistent with full employment equilibrium.

Now suppose that the real wage were to decline to w_C, so that the supply and effective demand for labor are equilibrated at point C. At point C, involuntary unemployment has vanished, but clearly this situation is not optimal. The reduced real wage has induced AB man-hours of labor to leave the labor force. Employment remains AB man-hours below the level associated with general equilibrium. Involuntary, i.e., excess supply, unemployment has been replaced by voluntary unemployment.[13]

The conclusion is that too high a real wage was not the cause of the lower employment, and a reduction in the real wage

[13] In terms of the *BLS* unemployment statistic, it is not clear that "zero" unemployment would be measured at w_C. If the higher wage, w^*, were (at least for a time) viewed as "normal," a considerable proportion of job seekers at wage w_C would be those willing to work at w^*, but not at w_C. These people are in the labor market seeking information on possible employment opportunities at (or above) w^*, and would not actually be willing to work at the going wage (see Armen Alchian). To the extent that the *BLS* measure includes this type of frustrated job seeker, the index will be a better measure of the gap between actual and general equilibrium employment BA, while simultaneously being a poorer index of those seeking but not obtaining employment at the going wage w_C.

is only a superficial cure. The real cause of the problem was the fall in commodity demand, and only a reflation of commodity demand can restore employment to the proper level.

The above analysis suggests the following cyclical patterns of real wages and employment: A decline in commodity demand and output produces a decline in employment with a corresponding excess supply of labor (point B). To the extent that real wages decline in response to this excess supply, a fall in real wages toward w_C will accompany (follow upon) the decline in employment. If, at point C or at some intermediate point between B and C, some action is taken to restore effective commodity demand, excess demand for labor (or, at least reduced excess supply) will result. In that case, a rising real wage may accompany the recovery of output and employment. Thus, disequilibrium analysis of the labor market suggests that real wages may move procyclically. This result differs from the conventional view that employment and real wages must be inversely related.

The present model can also be used to analyze involuntary unemployment which results from an excessive real wage. Clearly, if the real wage were above w^*, no stimulation of commodity demand could bring about full employment equilibrium, unless the real wage were reduced. This classical type of involuntary unemployment should be clearly distinguished from the type of unemployment discussed above, which arises, with the real wage at or below w^*, from a deficiency of demand for commodities.

IV. Clower's Analysis of the Consumption Function

In order to close the model, we must also analyze household behavior. Consider the behavior of the representative household under the provisional assumption that it regards utility maximization as being subject only to the budget constraint. In particular, the household perceives that it can sell all the labor which it supplies and purchase all the commodities which it demands at the existing levels of w and P. Assume the utility function to be

$$U = U\left(x^S, y^D, \frac{M}{P} + m^D\right),$$

with the partial derivatives $U_1<0$, $U_2>0$, and $U_3>0$. The budget constraint is

$$\pi + wx^S = y^D + m^D$$

x^S, y^D, and m^D may be denoted as the notional supply of labor, the notional demand for commodities, and the notional demand for additional money balances. Utility maximization in general will imply that x^S, y^D, and m^D are each functions of w, M/P, and π. For simplicity, we shall assume that x^S depends only on the real wage. The important point is that the notional demand functions for commodities and additional money balances do not have the forms of the usual consumption and saving functions with income as an argument, because the household simultaneously chooses the quantity of labor to sell.

Clower contrasts the above notional process to a situation in which labor services are in excess supply. Given voluntary exchange, actual total employment in this situation equals the total quantity demanded. Thus, the representative household is unable to sell its notional labor supply x^S and obtain its implied notional labor income wx^S.[14] Labor income is no longer a choice variable which is maximized out, but is instead exogenously given. We may assume that the representative household is able to obtain the quantity of employment x, where $x<x^S$,

[14] We assume that the household would actually like to sell x^S. As indicated in fn. 9, we assume for simplicity that excess demand for commodities never coexists with excess supply of labor.

so that its total income is $wx+\pi$. In this case, the utility maximization problem amounts to the optimal disposition of this income.

In other words, the household maximizes

$$U\left(x, y^{D'}, \frac{M}{P} + m^{D'}\right)$$

subject to $\pi+wx=y^{D'}+m^{D'}$. The variables $y^{D'}$ and $m^{D'}$ may be denoted as the effective demands for commodities and additional money balances. Utility maximization now implies

$$(2) \qquad y^{D'} = y^{D'}\left(\pi + wx, \frac{M}{P}\right),$$

and

$$(3) \qquad m^{D'} = m^{D'}\left(\pi + wx, \frac{M}{P}\right)$$

Note that, in aggregate, $\pi+wx=y=F(x)$. Thus, since all income accrues to the households, consumption and saving demand depend ultimately only on the level of employment and real money balances and not on the real wage rate. The constraint $x<x^S$ would generally imply $y^{D'}<y^D$ and $m^{D'}<m^D$, but as x approaches x^S, $y^{D'}$ and $m^{D'}$ approach y^D and m^D.[15]

The important property of equations (2) and (3) is that they do have the form of the usual Keynesian consumption and saving functions. Labor income enters the consumption and saving functions as it represents the constraint upon the demand for current output imposed by the excess supply of labor.

[15] To the extent that long-run employment (income) exceeds current employment (income), a household may be more willing to maintain a higher demand for commodities at the expense of money balances. In this case effective commodity demand would remain closer to notional demand, and the "income multiplier" (as depicted later in Figure 4) would be smaller. In general, the size of the effect of quantity constraints on effective demands will depend on whether the constraint is viewed as "permanent" or "transitory."

The formal analogy between the Clower and Patinkin models should be apparent from the derivations of equations (2), (3) and equation (1), or more particularly equations (1.1) and (1.2) in footnote 12. Patinkin's model involves profit maximization subject to an output constraint, whereas Clower's model involves utility maximization subject to an employment constraint.

V. General Disequilibrium Involving Excess Supply

In Patinkin's analysis, the effective demand for labor was derived for a given level of demand for current output. To close this model, the demand for current output must be explained. In Clower's analysis, the effective demand for current output was derived for a given level of demand for labor. To close this model, the demand for labor must be explained. Thus, the Patinkin and Clower analyses are essential complements. When appropriately joined, they form a complete picture of the determination of output and employment in a depressed economy.

Figure 2 depicts Clower's analysis of the commodity market. The notional supply schedule for commodities is a downward sloping function of the real wage. The two notional demand schedules are upward sloping functions, reflecting the effect of substitutability between consumption and leisure as well as a positive income effect. As the real wage rate rises, leisure becomes relatively more expensive, and households tend to work and consume more. The schedule corresponding to the general equilibrium price level P^* passes through the point A. At point A, which corresponds to point A in Figure 1, P^* and w^* are consistent with simultaneous notional equilibrium in both the labor and commodity markets. The other notional commodity demand schedule in Figure 2 corresponds to the higher price level P_1. Because of the real balance effect, this

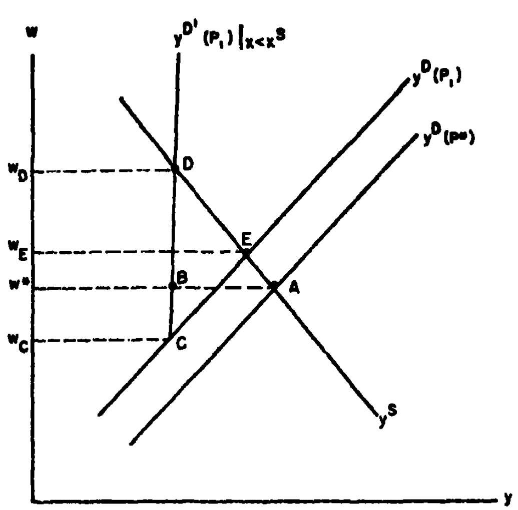


FIGURE 2. THE COMMODITY MARKET WITH EXCESS SUPPLY OF LABOR

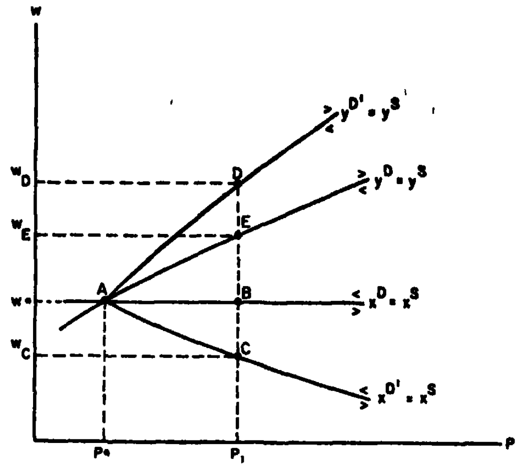


FIGURE 3. INTERACTION OF EXCESS SUPPLY IN BOTH MARKETS

curve lies to the left of the curve associated with P^*.[16] If $x=x^S$, the effective demand for commodities coincides with the notional demand. If $x<x^S$, the effective demand is independent of the real wage, as noted above, and is less than the notional demand. The effective demand schedule shown in Figure 2 corresponds to the higher price level P_1. Points B, C, D, and E also correspond to the same points in Figure 1. This correspondence can be seen most clearly by explicitly depicting the interaction between the two markets, as is done in Figures 3 and 4.

Figure 3 illustrates the relationship between the existence of excess supply in one market and the other. In Figure 3, the points A, B, C, D, and E coincide with the same points in Figures 1 and 2. The four loci separate the regions of inequality between the indicated supply and demand concepts. The locus $x^D=x^S$ is horizontal because, by assumption, both x^D and x^S depend only on the real wage. The locus $y^D=y^S$ is upward sloping because as shown in Figure 2, y^S is a decreasing function of the real wage, whereas y^D is an increasing function of the real wage (substitution and income effect) and a decreasing function of the price level (real balance effect). These loci intersect at point A, which depicts full employment general equilibrium. Points B, C, D, and E are all associated with a price level P_1, which is higher than the equilibrium price level P^*.[17] Point B, for example, would be consistent with notional equilibrium in the labor market, but implies excess supply in the commodity market. The essential point of Patinkin's analysis is that the effective demand for labor is smaller than the notional demand when commodities are in excess supply. Thus, the locus $x^{D'}=x^S$ exists to the right of point A and lies everywhere below the locus $x^D=x^S$. The existence of excess supply in the commodity market enlarges the region of excess supply in the labor market. Similarly, according to Clower's

[16] As the model is constructed, only y^D and m^D of the five notional schedules; x^D, x^S, y^D, y^S, and m^D depend on the price level independently of the real wage. In a more general model, real balances would affect x^D, x^S, and y^S, and the price level would affect these schedules also. By ignoring this possibility, the exposition is simplified without losing any of the essence of the analysis. Of course, if none of the five schedules were influenced by the price level, prices would not be determined within the model.

[17] We could, of course, just as well think of these points as being associated with a nominal money supply which is too small.

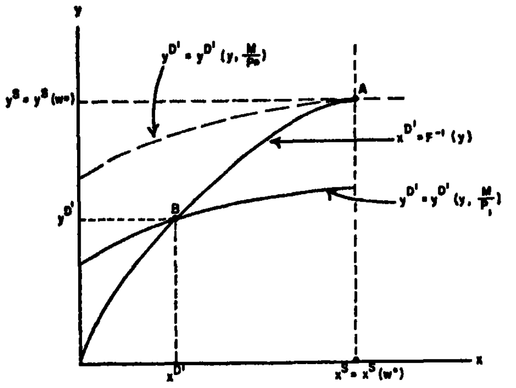


FIGURE 4. OUTPUT AND EMPLOYMENT WITH EXCESS SUPPLY IN BOTH MARKETS

analysis, the effective demand for commodities is less than the notional demand when labor is in excess supply. Thus, the locus $y^{D'}=y^S$ exists to the right of point A and lies everywhere above the locus $y^D=y^S$. The existence of excess supply in the labor market also enlarges the region of excess supply in the commodity market.

Figure 4 illustrates the determination of the actual quantities of current output and employment when there is excess supply in both markets. In particular, Figure 4 has been drawn under the assumption that the existing wage-price vector is (w^*, P_1), that is that the economy is at point B of Figures 1, 2, and 3. Given voluntary exchange, x and y are determined by $x=\min[x^{D'}, x^S]$ and $y=\min[y^{D'}, y^S]$. The solid locus $x^{D'}=F^{-1}(y)$ describes firm behavior for values of y less than y^S. The solid locus $y^{D'}=y^{D'}(y, M/P_1)$ describes household behavior for values of x less than x^S. The intersection of these two loci determines the values of x and y corresponding to point B. Point A, full employment equilibrium, is at the intersection of y^S and x^S. Since at point B the real wage is consistent with full employment equilibrium, a movement from B to A involves on net only a fall in the price level from P_1 to P^*. In Figure 4, this fall in P is represented by an upward shift in $y^{D'}$ to the dashed locus $y^{D'}(y, M/P^*)$, which intersects $x^{D'}$ at point A. The income multiplier in this case is given by the ratio of the difference between y^S and $y^{D'}(B)$ to the vertical distance between the two curves $y^{D'}(P^*)$ and $y^{D'}(P_1)$. Figure 4 is simply the Keynesian cross diagram with employment replacing income on the horizontal axis.

VI. General Disequilibrium Involving Excess Demand

The preceding discussion has concentrated on the case of excess supply in the markets for both commodities and labor. However, analogous considerations clearly apply to the boom situation of excess demand for both commodities and labor.

First, consider the behavior of the representative firm when there is excess demand for labor. The representative firm will be be able to obtain the quantity of labor x, where $x<x^D$. The firm then must maximize

$$\pi = y^{S'} - wx$$

subject to $y=F(x)$. The variable $y^{S'}$ may be denoted as the effective supply of commodities. The problem is simply to produce as much output as possible with the available labor. The solution is

$$(4) \qquad y^{S'} = F(x) \quad \text{for } \frac{dF}{dx} \geq w$$

Figure 5 depicts the commodity market in this situation, and is analogous to Figure 2. The price level P_2 is assumed to be below P^*.

Next, consider the behavior of the representative household when there is excess demand for commodities. The representative household will be able to obtain the quantity of commodities y, where $y<y^D$. The household then has to choose between either saving, i.e., accumulating as money balances the income which it cannot spend on consumption, or substituting leisure for the unobtainable commodities by supplying less labor, or some combination of the

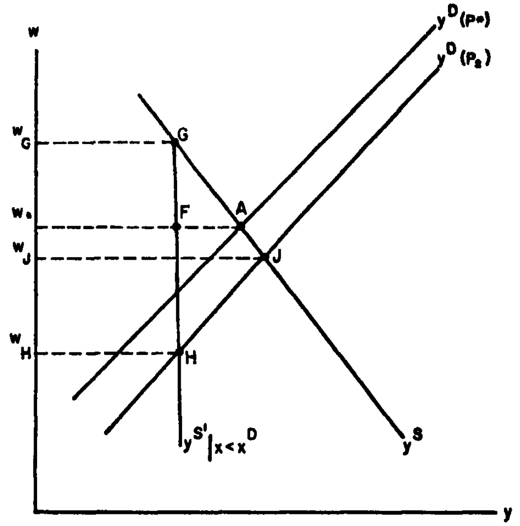


FIGURE 5. THE COMMODITY MARKET WITH EXCESS DEMAND FOR LABOR

two. Formally, the household's problem is to maximize

$$U\left(x^{S'}, y, \frac{M}{P} + m^{D'}\right)$$

$$\text{subject to} \qquad \pi + wx^{S'} = y + m^{D'}$$

The variable $x^{S'}$ may be denoted as the effective supply of labor. Utility maximization now implies

$$x^{S'} = x^{S'}\left(w, \frac{M}{P}, \pi, y\right), \tag{5}$$

and

$$m^{D'} = m^{D'}\left(w, \frac{M}{P}, \pi, y\right) \tag{6}$$

This theory stresses the fact that a household may react to frustrated commodity demand in two ways. First, the household may save the income which cannot be spent on consumption (in this model, solely by augmenting money balances). This option corresponds to the classical concept of forced saving, or, more precisely, what D. H. Robertson defined as "automatic lacking." Second, the household may increase leisure by reducing its supply of labor. The second option probably becomes more important when excess commodity demand is chronic, as in wartime or during other periods of rationing and price controls.[18] However, given that consumption, saving, and leisure in aggregate are substitutes, in general some combination of the two options will always be optimal. Excess demand will generally result in some fall in output.

Classical analysis, in which labor supply is solely a function of the real wage, assumes that households channel all frustrated commodity demand into forced saving. The possibility of reduced labor supply is ignored. However, the inclusion of this option is especially interesting, since it has the apparently paradoxical implication that excess commodity demand can result in decreased employment and output.

Figure 6, which is analogous to Figure 1, depicts the labor market in this situation. Two important observations should be stressed. First, too low a real wage, that is a real wage below the level consistent with general equilibrium, is not a necessary condition for excess demand for labor, even though the notional demand and supply for labor are both assumed to depend only upon the real wage. This observation is obviously the converse of the earlier observation that the effective demand for labor is not uniquely associated with the real wage. If commodities are in excess demand so that, given voluntary exchange, $y<y^D$, which in turn implies $x^{S'}<x^S$, at real wage w^* excess demand for labor will amount to quantity AF.

Second, with commodities in excess demand, the quantity of employment will generally be below the full employment level. The explanation of this apparent

[18] For example, R. Vicker, a recent visitor to the Soviet Union, reports the effects of suppressed inflation upon output: "Goods produced for sale in state retail outlets are snapped up more and more quickly, and the remaining excess of income over things to spend it on dilutes the incentive of Soviet workers."

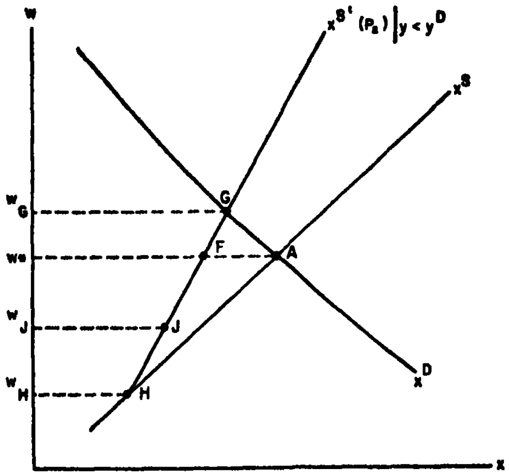


FIGURE 6. THE LABOR MARKET WITH EXCESS DEMAND FOR COMMODITIES

paradox, as indicated above, is twofold: 1) the quantity of employment can be no greater than the quantity supplied; and 2) when their consumption plans are frustrated households will generally substitute leisure and thus supply less labor at any given real wage. Notice that even if the real wage should rise sufficiently, i.e., to w_G, to eliminate the excess demand for labor, the level of employment would still be below that obtaining at general equilibrium.

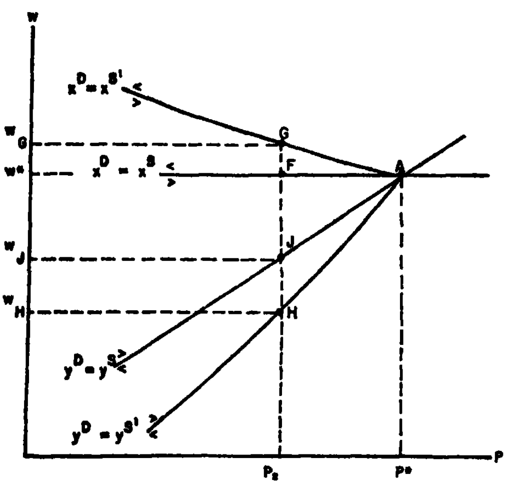


FIGURE 7. INTERACTION OF EXCESS DEMAND IN BOTH MARKETS

Finally, Figures 7 and 8, which are analogous to Figures 3 and 4, depict the interaction between the two markets with excess demand in both. Points A, F, G, H, and J in Figure 7 coincide with the same points in Figures 5 and 6. Figure 8 is drawn under the assumption that the existing wage-price vector is (w^*, P_2), that is, that the economy is at point F. The details of the construction of these diagrams are left as an exercise for the reader.

VII. Summary

This paper describes the application of a general disequilibrium approach to familiar problems of macro-analysis. Some familiar results, such as the notion that insufficient commodity demand produces unemployment, are arrived at in a much more satisfactory manner than is possible under more conventional analysis. In addition, the specific inclusion of disequilibrium elements leads to some nonfamiliar results.

The impact of excess supply of commodities on labor demand removes the one-to-one classical relationship between real wage and employment. In a general disequilibrium situation, unemployment can coexist with "non-excessive" real wages, and a procyclical pattern of real wages is consistent with the theoretical model.

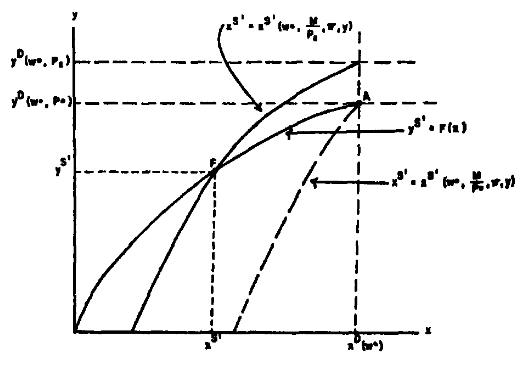


FIGURE 8. OUTPUT AND EMPLOYMENT WITH EXCESS DEMAND IN BOTH MARKETS

The disequilibrium analysis of the commodity market is formally parallel to the analysis of the labor market. The Keynesian consumption function emerges as a manifestation of the impact of excess labor supply on commodity demand. In this respect conventional macro-analysis is seen to be asymmetric. On the one hand, the disequilibrium impact of excess labor supply is implicitly recognized by entering income as a separate argument in the consumption function. However, on the other hand, the impact of excess commodity supply is neglected by adhering to the classical labor demand function which involves only the real wage. Because of this peculiar asymmetry, previous analyses of unemployment have had to rely on such contrived devices as a countercyclical pattern of real wages or fixed proportion production functions.

The framework for analyzing the excess supply, depression case is directly applicable to an analysis of sustained excess demand. The classical concept of forced saving is one aspect of the impact of excess commodity demand on household decision making. The forced saving solution is, however, incomplete, since labor supply would also react inversely to a prolonged frustration of commodity demand. To the extent that labor supply declines in response to excess commodity demand, increases in commodity demand lead to reduced employment, rather than to increased (forced) saving.

REFERENCES

A. A. Alchian, "Information Costs, Pricing, and Resource Unemployment," *Western Econ. J.*, June 1969, *7*, 109–28.

K. J. Arrow, "Toward a Theory of Price Adjustment," in M. Abramowitz, ed., *The Allocation of Economic Resources*, Stanford 1959.

R. J. Barro, "A Theory of Monopolistic Price Adjustment," read at Econometric Society Meetings, Detroit, Dec. 1970.

———, "A Theory of Optimal Adjustment," forthcoming 1971.

R. G. Bodkin, "Real Wages and Cyclical Variations in Employment," *Can. J. Econ.*, Aug. 1969, *2*, 353–74.

R. Clower, "The Keynesian Counter-Revolution: A Theoretical Appraisal," in F. H. Hahn and F. P. R. Brechling, eds., *The Theory of Interest Rates*, London 1965.

E. O. Edwards, "Classical and Keynesian Employment Theories: A Reconciliation," *Quart. J. Econ.*, Aug. 1959, *73*, 407–28.

H. I. Grossman, "Money, Interest, and Prices in Market Disequilibrium," *J. Polit. Econ.*, forthcoming 1971.

J. Hicks, *Capital and Growth*, New York, 1965.

J. M. Keynes, *The General Theory of Employment, Interest, and Money*, New York 1936.

———, "Relative Movements of Real Wages and Output," *Econ. J.*, Mar. 1939, *49*, 34–51.

E. Kuh, "Unemployment, Production Functions, and Effective Demand," *J. Polit. Econ.*, June 1966, *74*, 238–49.

A. Leijonhufvud, *On Keynesian Economics and the Economics of Keynes*, New York 1968.

D. Patinkin, "Involuntary Unemployment and the Keynesian Supply Function," *Econ. J.*, Sept. 1949, *59*, 360–83.

———, *Money, Interest, and Prices*, 1956: 2d ed. New York 1965.

D. H. Robertson, *Banking Policy and the Price Level*, London 1926.

R. M. Solow and J. E. Stiglitz, "Output, Employment, and Wages in the Short Run," *Quart. J. Econ.*, Nov. 1968, *82*, 537–60.

R. Vicker, "USSR: Rising Income, Black Markets," *The Wall Street Journal*, Apr. 21, 1970.

A Test for Relative Efficiency and Application to Indian Agriculture

By LAWRENCE J. LAU AND PAN A. YOTOPOULOS*

Economic efficiency is an elusive concept in which the economist, the engineer, and the policy maker all have great stakes.[1] The policy implications of economic efficiency permeate both the micro- and the macroeconomic level. Suppose, for example, that we can measure the efficiency of small and large farms. We can then determine by how much a given set of farms could be expected to increase its output through appropriate reorganization without absorbing additional resources in the aggregate.[2] We can also draw policy recommendations in connection with land ceilings, land redistribution, and land groupings under cooperative farming and other forms of agrarian organization.[3]

The difficulties with the existing approaches to efficiency are both conceptual and empirical. We will illustrate by pointing out the ambiguities of the conventional variants of efficiency, which we will classify as economic efficiency, price or allocative efficiency and technical efficiency.[4]

The simplest—and most naive—measure of economic efficiency is a partial productivity index, usually that of labor, although occasionally farm land (see Morton Paglin, 1965). This approach ignores the presence of other factors which affect average (and marginal) productivity.[5] A more sophisticated approach constructs indexes of efficiency that consist of a weighted average of inputs (the weights being either relative prices or relative factor shares) which is compared to output. Such an index is basically an output-cost ratio (see Paglin 1965, Robert Bennett). This approach runs into the usual index number problems that have been so aptly summarized by Evsey Domar.

Price or allocative efficiency traditionally rests on an index of marginal product and opportunity cost.[6] A number of problems arise in connection with this approach to price efficiency. First, it is an absolute concept which is of doubtful usefulness when one compares different groups of firms, even after allowing for differences in production functions and input prices. If among all inputs, the ratios of marginal products to opportunity costs are equal to one, a firm is price-efficient. Direct comparison among firms that satisfy

* The authors are, respectively, assistant professor, department of economics, and associate professor, Food Research Institute, Stanford University. An earlier draft of this paper benefited from the comments of A. S. Goldberger, B. F. Johnston, P. Zarembka, and the members of the Berkeley-Stanford Mathematical Economics and Econometrics Seminar. Mr. K. Somel provided able computational assistance. We also acknowledge financial support for Lau's participation by the National Science Foundation through Grant GS-2874 to Stanford University.

[1] For a discussion of the conceptual and theoretical complications that arise in connection with efficiency, see Margaret Lady Hall and Christopher Winsten.

[2] This is the approach of Harvey Leibenstein and William Comanor.

[3] Most frequently studies of efficiency lead to this kind of microeconomic policy implication. Specific examples are supplied below.

[4] In our own formulation below each of these terms will be given specific and rigorous conceptual and empirical content.

[5] These other factors can be ignored with impunity only under certain special circumstances.

[6] For examples, see Theodore Schultz, W. David Hopper, Yotopoulos (1968 a and b).

this equality to different degrees is almost impossible, as the literature on the second best has demonstrated. Second, it is a rather rigid concept that does not allow for possible differences in the initial endowment of fixed factors.

The conventional measurement of technical efficiency concentrates on the neutral displacement of the production function either between groups of firms or over time (see Irving Hoch, Yair Mundlak). Contrary to price efficiency which is purely a behavioral concept, technical efficiency is purely an engineering concept. It entirely abstracts from the effect of prices. An alternative approach to technical efficiency has been suggested by Michael Farrell. Under the assumption of constant returns to scale Farrell derives the "pessimistic" unit-isoquant, i.e., the isoquant which envelops the observations in the inputs-unit output space in such a way that no observation lies between the pessimistic isoquant and the origin. The index of efficiency is then constructed by measuring the deviation of a specific observation from the pessimistic isoquant. Besides also ignoring the effects of relative prices, the Farrell approach has the additional disadvantage that arises when one attempts to describe a stochastic universe by a deterministic process. The pessimistic isoquant is extremely sensitive to "outliers."[7]

The deficiencies of the existing approaches to measuring efficiency should dictate the minimum requirements that a new concept of relative *economic efficiency* should meet if it is to be at all useful. (i) It should account for firms that produce different quantities of output from a given set of measured inputs of production. This is the component of differences in *technical efficiency*. (ii) It should take into account that different firms succeed to varying degrees in maximizing profits, i.e., in equating the value of the marginal product of each variable factor of production to its price. This is the component of *price efficiency*. (iii) The test should take into account that firms operate at different sets of market prices. The decision rule on profit maximization yields actual profits (as well as quantity of output supplied and quantities of variable inputs demanded) as a function, *inter alia*, of input prices. It is clear that two firms of equal technical efficiency which have successfully maximized profits would still have different value of profits as long as they face different prices.

The interrelationships of the concepts of technical efficiency, price efficiency and economic efficiency can be explained in an intuitive way. Consider two firms with production functions identical up to a neutral displacement parameter,

$$V^1 = A^1F(X^1); \qquad V^2 = A^2F(X^2),$$

where V is the output, A the technical efficiency parameter, F the production function, X the vector of inputs employed, and the superscript denotes firm.

A firm is considered more technical-efficient than another, if, given the same quantities of measurable inputs, it consistently produces a larger output. Firm 1 is more technical-efficient than firm 2 if $A^1 > A^2$.

A firm is price-efficient if it maximizes profits, i.e., it equates the value of the marginal product of each variable input to its price. A firm which fails to maximize profits is, by definition, price inefficient. Consider now two complications in connection with the definition of price efficiency. First, assume that the prices of inputs are different for each firm. Firms now equate the value of the marginal product of each factor to its firm-specific opportunity cost. Second, firms may not maxi-

[7] The approach has been modified to partly account for this shortcoming by Dennis Aigner and S. F. Chu and C. Peter Timmer.

mize profits. For such firms the usual marginal conditions do not hold. It is assumed that these firms equate the value of the marginal product of each factor to a constant (which may be firm- and factor-specific) proportion of the respective firm-specific factor prices, i.e., and for firm 1,

$$p\frac{\partial V^1}{\partial X_i^1} = \overset{1}{k_i}\overset{1}{c_i}, \qquad \overset{1}{k_i} \geq 0$$

In this case k_i^1 indexes the decision rule that describes the firm's "profit-maximizing" behavior with respect to factor i. It encompasses perfect profit maximization as a special case when $k_i^1 = 1$ for all i. Now consider two price-inefficient firms of equal technical efficiency and facing identical output and input prices. The firm with the higher profits within a certain range of prices is considered the *relatively* more *price-efficient* firm (within the same range of prices).

Economic efficiency combines both technical and price efficiency. For this purpose consider two firms of *varying degrees of technical and price efficiency* but facing identical prices. The firm with the higher profits within a certain range of prices is considered the *relatively* more *economic-efficient* firm.

The concept of the profit function, as first introduced by D. L. McFadden, becomes operationally the ideal tool for our approach.[8] In Section I, we develop the theory of the profit function in its general form, without introducing firm-specific (and input-specific) price efficiency decision rules or firm-specific technical efficiency parameters. These are introduced in Section II which formulates the test of relative economic efficiency for the general case. In Section III we make the test of Section II operational by casting it in the framework of a Cobb-Douglas function.

Section IV is based on data from the *Farm Management Studies* (*The Studies*) of the Indian Ministry of Food and Agriculture. A number of researchers have used the same body of data to draw efficiency implications between small and large farms in India: A. K. Sen (1964, 1966); Paglin (1965); Bennett; G. S. Sahota; A. M. Khusro; to mention only a few. The findings are generally contradictory and inconclusive.[9] This comes as small surprise, given the divergent and often ambiguous concepts of efficiency that the authors have used.

A warning is in order at this point. We share the reservations of the previous authors about the limitations and the reliability of the data of *The Studies*. The reader, therefore, is urged to interpret cautiously our finding that small farms are economically more efficient than large farms. We intend our empirical application as an illustration of a method of measuring relative efficiency. It is a method that is based on the precepts of economic theory, it is more general than the existing alternatives, it is operational and it is parsimonious from the point of view of data requirements. Needless to say, the usefulness of our test is not restricted to agriculture nor is it specific for comparing small and large farms. Actually much more important insights into the form of economic organization might be forthcoming if one compares different groupings, such as owners versus share tenants, leaseholders versus tenants, adopters of new varieties versus nonadopters. Similar ramifications can be suggested for other fields of economics.

[8] Marc Nerlove first proposed a measurement of relative economic efficiency based on a profit function. However his approach is different from ours.

[9] For a summary of this discussion see Jagdish Bhagwati and Sukhamoy Chakravarty, and Yotopoulos, Lau, and Kutlu Somel.

I. The Profit Function

Consider a firm with a production function with the usual neoclassical properties

$$(1) \quad V = F(X_1, \ldots, X_m; Z_1, \ldots, Z_n)$$

where V is output, X_i represents variable inputs, and Z_i represents fixed inputs of production. Profit (defined as current revenues less current total variable costs) can be written

$$(2) \quad P' = pF(X_1, \ldots, X_m; Z_1, \ldots, Z_n) - \sum_{i=1}^{m} c_i' X_i$$

where P' is profit, p is the unit price of output, and c_i' is the unit price of the ith variable input. The fixed costs are ignored since, as it is well known, they do not affect the optimal combination of the variable inputs.

Assume that a firm maximizes profits given the levels of its technical efficiency and fixed inputs. The marginal productivity conditions for such a firm are

$$(3) \quad p\frac{\partial F(X;Z)}{\partial X_i} = c_i', \qquad i = 1, \ldots, m$$

By using the price of the output as numeraire we may define $c_i \equiv c_i'/p$ as the normalized price of the ith input. We can then write (3) as

$$(4) \quad \frac{\partial F}{\partial X_i} = c_i, \qquad i = 1, \ldots, m$$

By similar deflation by the price of output we can rewrite (2) as (5) where we define P as the "Unit-Output-Price" profit (or *UOP* profit)

$$(5) \quad P = \frac{P'}{p} = F(X_1, \ldots, X_m; Z_1, \ldots, Z_n) - \sum_{i=1}^{m} c_i X_i$$

Equation (4) may be solved for the optimal quantities of variable inputs, denoted X_i^*'s, as functions of the normalized prices of the variable inputs and of the quantities of the fixed inputs,[10]

$$(6) \quad X_i^* = f_i(c, Z), \qquad i = 1, \ldots, m$$

where c and Z are the vectors of normalized input prices and quantities of fixed inputs, respectively.

By substitution of (6) into (2) we get the *profit function*,[11]

$$(7) \quad \Pi = p\left[F(X_1^*, \ldots, X_m^*; Z_1, \ldots, Z_n) - \sum_{i=1}^{m} c_i X_i^*\right] = G(p, c_1', \ldots, c_m'; Z_1, \ldots, Z_n)$$

The profit function gives the *maximized* value of the profit for each set of values $\{p;\ c_1', \ldots, c_m';\ Z_1, \ldots, Z_n\}$. Observe that the term within square brackets on the right-hand side of (7) is a function only of c and Z. Hence we can write

$$(8) \quad \Pi = pG^*(c_1, \ldots, c_m; Z_1, \ldots, Z_n)$$

The *UOP* profit function is therefore given by

$$(9) \quad \Pi^* = \frac{\Pi}{p} = G^*(c_1, \ldots, c_m; Z_1, \ldots, Z_n)$$

Observe also that maximization of profit in (2) is equivalent to maximization of *UOP* profit in (5) in that they yield identical values for the optimal X_i^*'s. Hence Π^* in (9) indeed gives the maximized value of *UOP* profit in (5). We employ the *UOP* profit function Π^* because it is easier to work with than Π. It is evident that given Π^* one can always find Π, and vice versa.

[10] The unsubscripted variables X, Z, c', c, X^i, Z^i, c^i, and k^i are used to denote vectors. Superscripts, as above, denote firms.

[11] One should be careful to distinguish between profit as defined in (2) and the profit function in (7)

On the basis of a priori theoretical considerations we know that the *UOP* profit function is decreasing and convex in the normalized prices of variable inputs and increasing in quantities of fixed inputs. It follows also that the *UOP* profit function is increasing in the price of the output.

A set of dual transformation relations connects the production function and the profit function.[12] The most important one, from the point of view of our application here, is what is sometimes referred to as the Shephard-Uzawa-McFadden Lemma, as shown in equations (10) and (11).

$$(10) \quad X_i^* = -\frac{\partial \Pi^*(c, Z)}{\partial c_i}, \quad i = 1, \ldots, m,$$

$$(11) \quad V^* = \Pi^*(c, Z) - \sum_{i=1}^{m} \frac{\partial \Pi^*(c, Z)}{\partial c_i} c_i$$

where V^* is the supply function.

At this point we should emphasize the advantages of working with the *UOP* profit function instead of the traditional production function. First, the Shephard-Uzawa-McFadden Lemma allows us to derive the firm's supply function, V^*, and the firm's factor demand functions, X_i^*'s, directly from the *UOP* profit function of (9) instead of solving equation (4) which involves the production function.[13] Second, it is clear that the supply function and the factor demand functions may be obtained by simply starting with an *arbitrary UOP* profit function which is decreasing and convex in the normalized prices of the variable inputs and increasing in the fixed inputs. In addition, by duality, as McFadden has shown, there exists a one-to-one correspondence between the set of concave production functions and the set of convex profit functions.[14] Every concave production function has a dual which is a convex profit function, and vice versa.[15] Hence, without loss of generality, one can consider for profit-maximizing, price-taking firms, only profit functions in the analysis of their behavior *without* an explicit specification of the corresponding production function. This provides a great deal of flexibility in empirical analysis. Third, by starting from a profit function, we are assured by duality that the resulting system of supply and factor demand functions is obtainable from the maximization of a concave production function subject to given fixed inputs and under competitive markets. Fourth, the profit function, the supply function, and the derived demand functions so obtained are functions only of the normalized input prices and the quantities of fixed inputs, variables that are normally considered to be determined independently of the firm's behavior. Econometrically, this implies that these variables are exogenous variables, and by estimating these functions we avoid the problem of simultaneous equations bias to the extent that it is present.

II. Relative Economic Efficiency

The discussion of the profit function in Section I is general. It does not consider differences in technical efficiency and differences in price efficiency that might exist between firms. The purpose of this section is to introduce such differences and to combine them into the concept of rela-

[12] These relations are given and proven in McFadden and Lau.

[13] One practical advantage of using *UOP* profit functions as opposed to deriving the factor demand equations directly from equation (4) is that in many cases equation (4) cannot be solved in closed form.

[14] There are additional regularity conditions on the production and profit functions which are spelled out in detail in McFadden. Since we are interested in the empirical application of profit functions, we will not be concerned with the finer details. It suffices to say that almost all continuous production functions in current use which are concave will give rise to a well-behaved profit function.

[15] We rule out constant returns to scale in the variable factors, which, as is well known, would lead to indeterminate output and input levels. See Lau.

tive economic efficiency. Our approach is straightforward. Given comparable endowments, identical technology, and normalized input prices, the *UOP* profits of two firms should be identical if they have both maximized profits. To the extent that the one firm is more price-efficient, or more technically efficient, than the other, the *UOP* profits will differ even for the same normalized input prices and endowments of fixed inputs.

Let us represent the situation as follows. For each of two firms the production function is given by

$$
\begin{aligned}
V^1 &= A^1F(X^1, Z^1); \\
V^2 &= A^2F(X^2, Z^2)
\end{aligned}
\tag{12}
$$

where superscripts identify firms. The marginal conditions are given by

$$
\begin{aligned}
\frac{\partial A^1F(X^1, Z^1)}{\partial X_j^1} &= \overset{1}{k_j}\overset{1}{c_j} \\
\frac{\partial A^2F(X^2, Z^2)}{\partial X_j^2} &= \overset{2}{k_j}\overset{2}{c_j}
\end{aligned}
\tag{13}
$$

$$\overset{1}{k_j} \geqq 0,\ \overset{2}{k_j} \geqq 0,\ j = 1, \ldots, m$$

At this point it is useful to reiterate the basic differences in approach that equations (12) and (13) introduce, as compared to Section I. The formulation of Section I was general while now it becomes firm-specific. We can talk about relative efficiency only by comparing two or more firms. We allow for neutral differences in the production functions in terms of the firm-specific technical efficiency parameters, A^1 and A^2. They represent differences in environmental factors, in managerial ability and in other nonmeasurable fixed factors of production. If the two firms are equally technical-efficient, $A^1=A^2$. Furthermore, we now allow for a firm to be unsuccessful in its attempts to equate values of the marginal products of its inputs to their respective normalized prices. This is introduced through the firm-specific and variable input-specific k's.[16] If, and only if, two firms are equally price-efficient with respect to all variable inputs, then $k_i^1=k_i^2$, $i=1, \ldots, m$. We have defined economic efficiency to encompass both technical and price efficiency. In terms of our notation, therefore, the null hypothesis of equal relative economic efficiency for firm 1 and firm 2 implies that $A^1=A^2$ and $k^1=k^2$. The purpose of this section, therefore, is to develop a method to enable us to make this comparison.

In our formulation, the k's reflect a general systematic rule of behavior—a decision rule that gives the profit-maximizing marginal productivity conditions as a special case. That the decision rule for the firm consists of equating the marginal product to a constant times the normalized price of each input may be rationalized as follows: i) Consistent over- or under-valuation of the opportunity costs of the resources by the firm; ii) Satisficing behavior; iii) Divergence of expected and actual normalized prices; iv) Divergence of the subjective probability distribution of the normalized prices from the objective distribution of normalized prices; v) The elements of k^i may be interpreted as the first-order coefficients of a Taylor's series expansion of arbitrary decision rules of the type

$$\frac{\partial F}{\partial X_j^i} = f_j^i(c_j^i), \qquad i = 1, 2; j = 1, \ldots, m$$

where $f_j^i(0)=0$ and $f_j^{i\prime}(c_j^i) \geqq 0$. A wide class of decision rules may be encompassed under v). Observe that the right-hand sides of equation (13) may be interpreted as the "effective" prices facing the two firms. The behavior of the two firms can

[16] Of course, if a firm is perfectly successful in equalizing the normalized price of an input i to its opportunity cost, k_i assumes the value of one for that specific input.

then be viewed as profit-maximization subject to these effective prices and can be represented by the *behavioral UOP* profit function.

Let $G^*(c, Z)$ be the *UOP* profit function corresponding to $F(X, Z)$. By a well-known theorem proved in McFadden, the *UOP* profit function corresponding to a production function

$$V = AF(X, Z) \quad \text{is}$$

$$\Pi^* = AG^*(c/A, Z) \tag{14}$$

Recall that the $k_j^i c_j^i$'s may be interpreted as the effective prices. Thus we may write for the behavioral *UOP* profit functions of the two firms, respectively,

$$\begin{aligned} \Pi_b^1 &= A^1 G^*(k_1^1 c_1^1/A^1, \ldots, k_m^1 c_m^1/A^1; \\ &\qquad Z_1^1, \ldots, Z_n^1) \\ \Pi_b^2 &= A^2 G^*(k_1^2 c_1^2/A^2, \ldots, k_m^2 c_m^2/A^2; \\ &\qquad Z_1^2, \ldots, Z_n^2) \end{aligned} \tag{15}$$

As in the previous section, the demand functions are given by the Shephard-Uzawa-McFadden Lemma. We now, however, differentiate the behavioral *UOP* profit functions with respect to the effective prices $k_j^1 c_j^1$'s and $k_j^2 c_j^2$'s. We write[17]

$$\begin{aligned} X_j^i &= -A^i \frac{\partial G^*(k^i c^i/A^i; Z^i)}{\partial k_j^i c_j^i} \\ &= \frac{-A^i}{k_j^i} \frac{\partial G^*(k^i c^i/A^i; Z^i)}{\partial c_j^i}, \\ &\qquad i = 1, 2; j = 1, \ldots, m \end{aligned} \tag{16}$$

By correspondence from (11) the supply functions are now given by

$$\begin{aligned} V^i &= A^i G^*(k^i c^i/A^i; Z^i) \\ &\quad - A^i \sum_{j=1}^m k_j^i c_j^i \frac{\partial G^*(k^i c^i/A^i; Z^i)}{\partial k_j^i c_j^i} \\ &= A^i G^*(k^i c^i/A^i; Z^i) \\ &\quad - A^i \sum_{j=1}^m c_j^i \frac{\partial G^*(k^i c^i/A^i; Z^i)}{\partial c_j^i}, \\ &\qquad i = 1, 2 \end{aligned} \tag{17}$$

It should be emphasized at this point that X_j^i and V^i as given in (16) and (17) are the actual quantities of inputs demanded and output supplied by firm i given the firm-specific A^i and k^i. When appropriate functional forms are specified for G, statistical tests can be devised to test the null hypothesis of equal economic efficiency, i.e., $A^1 = A^2$ and $k^1 = k^2$, although not all of the parameters may be independently identified and estimated.

An alternative approach to looking at the demand and supply functions is to examine the *actual UOP* profit function. From (16) and (17) we can obtain the actual *UOP* profit functions by using equation (5),

$$\begin{aligned} \Pi_a^i &= V^i - \sum_{j=1}^m c_j^i X_j^i \\ &= A^i G^*(k^i c^i/A^i; Z^i) \\ &\quad + A^i \sum_{j=1}^m \frac{(1 - k_j^i) c_j^i}{k_j^i} \\ &\quad \cdot \frac{\partial G^*(k^i c^i/A^i; Z^i)}{\partial c_j^i}, \qquad i = 1, 2 \end{aligned} \tag{18}$$

Observe that i) $\partial \pi_a^i / \partial A^i > 0$, i.e., actual profit always increases with the level of technical efficiency for given normalized input prices and k^i; ii) When $k_j^i = 1$ for $j = 1, \ldots, m$, i.e., the firm is a true profit maximizer, the actual and behavioral *UOP* profit functions coincide; iii) When $A^1 = A^2$ and $k^1 = k^2$, the actual *UOP* functions of the two firms coincide with each other. Therefore one can also test the null hypothesis of equal relative economic efficiency by comparing the actual *UOP* profit functions of the two firms when appropriate functional forms are specified

[17] To simplify notation we omitted the asterisks from the demand and supply functions.

for G. This is the approach that will be employed in our empirical analysis.[18]

An additional test becomes relevant if we reject the joint hypothesis that $(A^1, k^1) = (A^2, k^2)$. In this case an overall indication of the relative efficiency between the two firms within a specified range of normalized prices for variable inputs may be obtained by comparing the actual values of the *UOP* profit functions within this range. If $\Pi_a^1 \geq \Pi_a^2$ for all normalized prices within a specified range, then clearly, the first firm is relatively more efficient within the price range. If some knowledge on the probability distribution of the future prices is available, a choice may be made as to the relative efficiency of the two firms.

One can also test the hypothesis that the fixed inputs command equal rent on the two firms by computing the first derivatives of the actual *UOP* profit functions with respect to the fixed inputs and testing for their equality. This may have important implications for the optimal form of economic organization in terms of the distribution of fixed inputs.

III. The Formulation of the Cobb-Douglas Case

In this section we proceed to specify the appropriate functional form of the profit function and formulate empirically the test of relative economic efficiency. For this purpose one can start from a Cobb-Douglas, or for that matter, from any other form of a function. We cast our analysis in terms of the Cobb-Douglas function because it appears superior through tests of alternative functional forms.[19]

A Cobb-Douglas production function with decreasing returns in the m variable inputs and with n fixed inputs is given by[20]

$$V = A\left(\prod_{i=1}^{m} X_i^{\alpha_i}\right)\left(\prod_{j=1}^{n} Z_j^{\beta_j}\right)$$

where

$$\mu = \sum_{i=1}^{m} \alpha_i < 1$$

The *UOP* profit function is given by

$$\Pi^* = A^{(1-\mu)^{-1}}(1-\mu) \cdot \left(\prod_{i=1}^{m} (c_i/\alpha_i)^{-\alpha_i(1-\mu)^{-1}}\right) \cdot \left(\prod_{j=1}^{n} Z_j^{\beta_j(1-\mu)^{-1}}\right) \tag{19}$$

By direct computation, the actual *UOP* profit functions and the demand functions for this Cobb-Douglas production function are given in equations (20) and (21). It is clear that the actual *UOP* profit functions of the two firms differ by a constant factor, which is a function of the k_j^i's and A^i's. In addition, all the demand functions differ by constant factors. A test of equal economic efficiency will be based on the

[18] Note that by the profit identity, one of the system of profit, supply and demand functions is redundant and should be ignored in the actual estimation of the system. Otherwise the system variance-covariance matrix will be singular.

[19] These tests are presented in Yotopoulos, Lau, and Somel.

[20] The value of $\mu<1$ is required since constant or increasing returns in the variable inputs are inconsistent with profit maximization.

$$\Pi_a^i = \left[(A^i)^{(1-\mu)^{-1}}\left(1 - \sum_{j=1}^{m} \alpha_j/k_j^i\right)\right]\left[\prod_{j=1}^{m} (k_j^i)^{-\alpha_j(1-\mu)^{-1}}\right]\left[\prod_{j=1}^{m} \alpha_j^{-\alpha_j(1-\mu)^{-1}}\right] \cdot \left[\prod_{j=1}^{m} (c_j^i)^{-\alpha_j(1-\mu)^{-1}}\right]\left[\prod_{j=1}^{n} (Z_j^i)^{\beta_j(1-\mu)^{-1}}\right], \quad i = 1, 2 \tag{20}$$

$$X_l^i = (A^i)^{(1-\mu)^{-1}} (\alpha_l / k_l^i c_l^i) \left[\prod_{j=1}^{m} (k_j^i)^{-\alpha_j (1-\mu)^{-1}} \right] \left[\prod_{j=1}^{m} \alpha_j^{-\alpha_j (1-\mu)^{-1}} \right] \tag{21}$$
$$\cdot \left[\prod_{j=1}^{m} (c_j^i)^{-\alpha_j (1-\mu)^{-1}} \right] \left[\prod_{j=1}^{n} (Z_j^i)^{\beta_j (1-\mu)^{-1}} \right], \qquad i = 1, 2; \quad l = 1, \ldots, m$$

---

null hypothesis that all the constant factors of difference are ones.

Observe that the terms in the first three brackets of equation (20) involve constants. We thus define equation (22).

$$A_*^i \equiv (A^i)^{(1-\mu)^{-1}} \left(1 - \sum_{j=1}^{m} \alpha_j / k_j^i \right) \tag{22}$$
$$\left[\prod_{j=1}^{m} (k_j^i)^{-\alpha_j (1-\mu)^{-1}} \right] \left[\prod_{j=1}^{m} \alpha_j^{-\alpha_j (1-\mu)^{-1}} \right], \qquad i = 1, 2$$

Then the actual *UOP* profit functions are given by

$$\Pi_a^i = (A_*^i) \left[\prod_{j=1}^{m} (c_j^i)^{-\alpha_j (1-\mu)^{-1}} \right] \tag{23}$$
$$\cdot \left[\prod_{j=1}^{n} (Z_j^i)^{-\beta_j (1-\mu)^{-1}} \right], \qquad i = 1, 2$$

By writing A_*^2 and A_*^1 for firm 2 and firm 1, respectively, and taking the ratio of the constant terms we have

$$\frac{A_*^2}{A_*^1} = \left[\frac{A^2}{A^1} \right]^{(1-\mu)^{-1}} \frac{\left(1 - \sum_{j=1}^{m} \alpha_j / k_j^2 \right)}{\left(1 - \sum_{j=1}^{m} \alpha_j / k_j^1 \right)} \tag{24}$$
$$\cdot \left[\prod_{j=1}^{m} \left[\frac{k_j^2}{k_j^1} \right]^{-\alpha_j (1-\mu)^{-1}} \right]$$

Thus one may write, from equation (20)

$$\Pi_a^1 = A_*^1 \left[\prod_{j=1}^{m} (c_j^1)^{-\alpha_j (1-\mu)^{-1}} \right] \tag{25}$$
$$\cdot \left[\prod_{j=1}^{n} (Z_j^1)^{\beta_j (1-\mu)^{-1}} \right]$$

$$\Pi_a^2 = A_*^1 (A_*^2 / A_*^1) \left[\prod_{j=1}^{m} (c_j^2)^{-\alpha_j (1-\mu)^{-1}} \right] \tag{26}$$
$$\cdot \left[\prod_{j=1}^{n} (Z_j^2)^{\beta_j (1-\mu)^{-1}} \right]$$

Further defining

$$\alpha_j^* \equiv -\alpha_j (1-\mu)^{-1}; \tag{27}$$

and

$$\beta_j^* \equiv \beta_j (1-\mu)^{-1} \tag{28}$$

and taking natural logarithms of equations (25) and (26), we have

$$\ln \Pi_a^1 = \ln A_*^1 + \sum_{j=1}^{m} \alpha_j^* \ln c_j^1 + \sum_{j=1}^{n} \beta_j^* \ln Z_j^1, \tag{29}$$

$$\ln \Pi_a^2 = \ln A_*^1 + \ln \frac{A_*^2}{A_*^1} + \sum_{j=1}^{m} \alpha_j^* \ln c_j^2 + \sum_{j=1}^{n} \beta_j^* \ln Z_j^2 \tag{30}$$

We note that if $A^1 = A^2$ *and* $k^1 = k^2$, then $A_*^1 = A_*^2$ and the two functions Π_a^1 and Π_a^2 ($\ln \Pi_a^1$ and $\ln \Pi_a^2$) should be identical. This implies that $\ln A_*^2 / A_*^1 = 0$. We can

therefore test the equal relative efficiency hypothesis by utilizing a firm dummy variable in the logarithmic *UOP* profit function and examining if its value is equal to zero. It should be noted that for the Cobb-Douglas production function case, differences in technical efficiency and relative differences in price efficiency cannot be separately identified from the actual *UOP* profit functions.

IV. Empirical Implementation and Statistical Results

In this section we use data from *The Studies* of the Indian Ministry of Food and Agriculture to estimate the *UOP* profit functions for the small and large farms and to apply the test of equal economic efficiency for the two groups. *The Studies*, which have proven to be a bountiful data source for many researchers,[21] are based on cost-accounting records of 2,962 holdings in the six states of India and cover the three-year period, 1955–57.[22] All the data are, however, reported only in terms of averages of farms of a given size for each state. From the available raw data we proceed to specify as follows the variables of our analysis.

Output is given in terms of revenue V per farm in rupees; land T represents cultivable land per farm in acres,[23] and capital K is defined in terms of interest charges paid or imputed on the quantity of fixed capital per farm.[24] Labor is given in terms of labor days employed per farm as well as in terms of a labor cost per farm concept (i.e., cost of hired labor plus imputed cost of family labor). By dividing the latter labor concept through by the former we define the money wage rate per day, w'. Only three inputs are distinguished: labor, capital, and land. We treat labor as the variable input of production and land and capital as fixed inputs.[25] It appears reasonable, from both institutional reasons and from the periodic nature of the agricultural technology, that the latter may be considered as fixed inputs in the short run. Finally, from the revenue we subtract the total cost of variable inputs per farm, i.e., the wage bill, in order to define the profit variable, Π. It should be recalled that in the *UOP* profit function formulation of the preceding sections both Π^* and w are expressed in real terms. Unavailability of the prices for deflation poses a problem that will be discussed below.

For the Cobb-Douglas case, the profit function is given by (29) as

$$(31) \qquad \ln \Pi_a^1 = \ln A_*^1 + \alpha_1^* \ln w + \beta_1^* \ln K + \beta_2^* \ln T$$

[21] Besides the literature based on *The Studies* that is surveyed in Bhagwati and Chakravarty (especially pp. 40 ff.), one should also notice Paglin, Sahota and Yotopoulos, Lau and Somel.

[22] For this analysis we utilize data from the following states and years: West Bengal, Madras, Uttar Pradesh, Punjab, 1955–56; Madhya Pradesh, 1956–57. The latter is chosen because the 1955–56 report of *The Studies* for Madhya Pradesh does not contain comparable information as the others.

[23] It is assumed that the land input is homogeneous at least within states across farm sizes. This hypothesis was tested in Yotopoulos, Lau, and Somel.

[24] This definition of the capital concept is especially disturbing. Inasmuch as the interest rate used in the imputation—3 percent—is uniform throughout the states and the years, the true quantity of fixed capital will be proportional to our measure. This implicitly assumes that the flow of capital services as a ratio of the stock of capital is constant across farms. Such assumptions, as demonstrated by Yotopoulos (1967, 1968a), may lead to unreliable estimates of the coefficient of capital in a production function formulation. It appears that this may be the case with our estimated capital coefficient.

[25] Total other costs (i.e., costs other than labor costs, interest on fixed capital and land rent) should also be treated as a variable input of production. This is impossible in our profit function formulation due to the fact that we lack the "price" of other costs which is necessary for the *UOP* profit function. To the extent that the price of other costs varies only across states, its effect is captured by the state dummies. An alternative rationalization is that the other costs are employed in fixed proportions to output.

$$(32)\quad \ln \Pi_a^2 = \ln A_*^1 + \ln (A_*^2/A_*^1) + \alpha_1^* \ln w + \beta_1^* \ln K + \beta_2^* \ln T$$

where Π_a^i is actual *UOP* profit (total revenue less total variable cost, divided by the price of output), w is normalized wage rate, K is interest on fixed capital, and T is cultivable land. A maintained hypothesis is that the production function is identical on large and small farms up to a neutral efficiency parameter. This implies that the coefficients corresponding to $\ln w$, $\ln K$, and $\ln T$ are identical for large and small farms. A problem arises at this point. Our formulation of the *UOP* profit function is in terms of normalized input prices. However, in our empirical application these normalized input prices are not available since the data on money prices of output are rather poor. Fortunately, we note that equations (31) and (32) may be rewritten

$$\begin{aligned} \ln \Pi_a^1 &= \ln \Pi'^1 - \ln p \\ &= \ln A_*^1 + \alpha_1^* \ln w' - \alpha_1^* \ln p \\ &\quad + \beta_1^* \ln K + \beta_2^* \ln T \end{aligned}$$

or

$$\begin{aligned} \ln \Pi'^1 &= \ln A_*^1 + (1 - \alpha_1^*) \ln p \\ &\quad + \alpha_1^* \ln w' + \beta_1^* \ln K \\ &\quad + \beta_2^* \ln T \\ \ln \Pi'^2 &= \ln A_*^1 + (1 - \alpha_1^*) \ln p \\ &\quad + \ln \left(\frac{A_*^2}{A_*^1}\right) + \alpha_1^* \ln w' \\ &\quad + \beta_1^* \ln K + \beta_2^* \ln T \end{aligned}$$

where Π'^i is money profit in rupees, w', is the money wage rate in rupees per day, and p is the price of the output in rupees.

If the prices of outputs differ only across states, then one can insert state dummy variables to capture the effect of differences due to $(\ln A_* + (1-\alpha_1^*)\ln p)$. This also allows for interstate differences in the efficiency parameter in A_*. Hence our final estimating equation consists of

$$(33)\quad \begin{aligned} \ln \Pi &= \alpha_0 + S + \sum_{i=1}^{4} \delta_i^* D_i + \alpha_1^* \ln w' \\ &\quad + \beta_1^* \ln K + \beta_2^* \ln T \end{aligned}$$

where Π is farm profit in rupees (excluding interest on capital and land rent), w' is money wage rate, K is interest on fixed capital, T is cultivable land, the D_i's are state dummy variables and S is a dummy variable with value 1 for large farms and 0 for small farms. Large farms are defined as those with cultivable land greater than ten acres per farm.

A remark about the stochastic specification of the model is appropriate at this point. Not much is known about how disturbance terms in general should be introduced into economic relationships although Hoch, Mundlak and Hoch, and subsequently Zellner, Kmenta and Drèze have proposed one possible assumption that is workable in the Cobb-Douglas case. Here we assume that the error in the profits is due to climatic variations, divergence of the expected output price from the realized output price, imperfect knowledge of the technical efficiency parameter of the farm, and differences in technical efficiency among farms within the same size class. The demand functions are exact, or, in any case, if they are subject to error, the errors are uncorrelated with the errors of the logarithmic profit function, which are assumed to have zero expectation. Hence one can estimate the natural logarithms of the profit function alone with the least squares estimator, which in this case turns out to be minimum

variance, linear and unbiased. This specification is similar to the one used by Marc Nerlove in his pioneering study of empirical cost functions.

The results of the estimation are presented in Table 1. The F-value indicates that the hypothesis that all coefficients other than α_0 are zeroes should be rejected. The coefficient of the wage rate is negative while the coefficient of land is positive, in accord with a priori economic theory: the UOP profit function is decreasing in w' and increasing in T. The negative coefficient of capital can only be attributed to the misspecification of this variable that is due to the implicit assumption of proportionality between capital service flow and capital stock (Yotopoulos, 1967, 1968a). In addition, the second derivative of the UOP profit function with respect to the wage rate is

$$\frac{\partial^2\Pi^*}{\partial w^2} = \frac{\alpha_1^{*2}\Pi^*}{w^2} - \frac{\alpha_1^*\Pi^*}{w^2}$$

$$= \alpha_1^*(\alpha_1^* - 1)\frac{\Pi^*}{w^2}$$

$$= -2.141\,(-3.141)\frac{\Pi^*}{w^2} \geqq 0$$

as $\Pi > 0$ and hence $\Pi^* > 0$ for our whole sample. This also confirms the convexity assumption of the UOP profit function.

The estimates of Table 1 imply, by (27) and (28), estimates of the input elasticities of the production function. These are presented in Table 2. The elasticities appear reasonable by comparison with other available estimates of Cobb-Douglas agricultural production functions for India and other parts of the world.

We note that the capital coefficient as estimated directly also has a negative sign. Finally, the sum of elasticities obtained from the indirect estimates is somewhat larger than that obtained from the direct

TABLE 1—COBB-DOUGLAS PROFIT FUNCTION AND RELATED STATISTICS

Parameter	All Farms ($n=34$)
α_0	4.582
	(0.548)
S	−0.567
	(0.253)
δ_1^*	1.614*
	(0.549)
δ_2^*	−1.359*
	(1.274)
δ_3^*	−0.588*
	(0.485)
δ_4^*	0.296
	(0.715)
α_1^*	−2.141**
	(1.200)
β_1^*	−0.588
	(0.274)
β_2^*	1.797
	(0.233)
$\hat{\sigma}^2$	0.185
$\bar{R}^2$	0.896
F-Statistic	36.4

Source: Farm Management Studies

Notes: The estimating equation is

$$\ln \Pi = \alpha_0 + S + \sum_{i=1}^{4} \delta_i^* D_i + \alpha_1^* \ln w' + \beta_1^* \ln K + \beta_2^* \ln T$$

where

Π = profit, i.e., total revenue less total variable costs

w' = the money wage rate

S = dummy variable for farm size with value of one for large farms (greater than ten acres) and zero for small farms (less than ten acres)

D_i = regional dummy variable with D_1, D_2, D_3, D_4 taking the value of one for West Bengal, Madras, Madhya Pradesh, and Uttar Pradesh and zero elsewhere, respectively

K = interest on fixed capital

T = cultivable land in acres

$\hat{\sigma}^2$ = the estimate of the variance of the error in the equation

* Starred coefficients are not significantly different from zero at a probability level $\geqq$ 95 percent

** Double-starred coefficients are not significantly different from zero at a probability level $\geqq$ 95 percent; but they are significantly different from zero at a probability level $\geqq$ 90 percent.

All other coefficients are significantly different from zero at a probability level $\geqq$ 95 percent.

Two-tail test applies to the coefficients of the dummy variables; one-tail test to all other variables.

The standard errors of the estimated parameters are given in parentheses.

TABLE 2—COMPARISON OF DIRECT AND INDIRECT ESTIMATES OF INPUT ELASTICITIES OF THE PRODUCTION FUNCTION

Parameters	Direct Estimates	Indirect Estimates
α_1	.606	.682
β_1	−.103	−.187
β_2	.365	.572
$(\alpha_1+\beta_1+\beta_2)$	0.868	1.067

Source: Farm Management Studies.
Notes: The direct estimates are obtained by ordinary least squares regression of the natural logarithm of output on the natural logarithms of the three inputs, the farm size dummy, the four state dummies and the constant term. The indirect estimates of the parameters are derived from equations (27) and (28). Other notations as in Table 1.

estimates. It is slightly larger than one.[26] One may also add that the estimates obtained by fitting the profit function are statistically consistent, as opposed to those obtained directly from the production function by ordinary least squares, which are in general inconsistent because of the existence of simultaneous equations bias.

As we have indicated in the previous section the hypothesis of relative efficiency can be cast in terms of the constant term by which the two profit functions, one for small and one for large farms, differ. The null hypothesis is that the constant factor is equal to one. Furthermore, if one takes natural logarithms before estimating the profit function, the constant term becomes the coefficient of a dummy variable that differentiates the two groups of farms and the test becomes that the coefficient of the dummy variable is not significantly different from zero. Our results, therefore, reject the hypothesis of equal efficiency between the two groups. Furthermore, the sign of the dummy variable indicates that small farms are more profitable, i.e., more efficient, at all observed prices of the variable input, given the distribution of the fixed factors of production.

Given the actual profit functions for the two groups of farms, one can estimate the rate of return on the fixed inputs by computing the partial derivatives of the *UOP* profit function with respect to both capital and land. Suppressing interstate differences, one has

$$\frac{\partial \Pi}{\partial K} = \beta_1^* \frac{\Pi}{K} = -0.588\frac{\Pi}{K} \tag{34}$$

$$\frac{\partial \Pi}{\partial T} = \beta_2^* \frac{\Pi}{T} = 1.797 \frac{\Pi}{T} \tag{35}$$

These rates of returns are computed at the geometric mean of the large and small farms separately, and reported in Table 3. It is seen that both the rates of return on fixed capital and on land are larger on the small farms at the existing set of normalized prices faced by these farms.

V. Summary and Conclusion

In our formulation of the test of equal relative economic efficiency we use McFadden's profit function, which expresses a firm's maximized profit as a function of the prices of output and variable inputs of production and of the quantities of the

TABLE 3—COMPARISON OF THE RATES OF RETURN ON FIXED CAPITAL AND LAND BETWEEN LARGE AND SMALL FARMS

	Large Farms	Small Farms
Geometric Means		
$\bar{\Pi}$ (rupees)	2,184.62	493.90
$\bar{K}$ (rupees)	51.35	22.89
$\bar{T}$ (acres)	23.81	3.99
Rates of Return		
$\frac{\partial \Pi}{\partial K}$ (rupees per rupee)	−25.02	−12.69
$\frac{\partial \Pi}{\partial T}$ (rupees per acre)	164.88	222.44

Source: Farm Management Studies.

[26] An exact linear restriction is available for the testing of the hypothesis of constant returns to scale within the profit function framework. See Lau, also Lau and Yotopoulos.

fixed factors. In the Cobb-Douglas formulation the comparison of relative efficiency of two groups of firms is simply made by examining the coefficient of the group dummy variable.

A crucial feature of the profit function analysis is that it assumes firms behave according to certain decision rules, which include the profit maximization rules, given the price regime for output and variable inputs and given the quantities of their fixed factors of production. For the purposes of this paper, the existence of these systematic decision rules is a maintained hypothesis. It can, however, be tested directly within the framework developed by Wise and Yotopoulos.

The conclusion of the test of relative economic efficiency is in favor of the small farms (i.e., farms of less than ten acres). It appears that, given the fixed factors of production (land and fixed capital) and within the ranges of the observed prices of output and variable inputs (labor), the small farms of the sample of *The Studies* have higher actual profits. In the context of our analysis, this finding means that the small farms attain higher levels of price efficiency (i.e., of optimal price behavior) and/or they operate at higher levels of technical efficiency. This finding, should it be confirmed by similar tests with other sets of data, may imply that in agriculture the supervisory role of the owner-manager of the farm may be crucial for attaining high levels of economic efficiency. Within the context of this hypothesis, the test would draw the limits of the supervisory capacity of the manager at the ten acre farms.

REFERENCES

D. J. Aigner and S. F. Chu, "On Estimating the Industry Production Function," *Amer. Econ. Rev.*, Sept. 1968, *58*, 826–39.

R. L. Bennett, "Surplus Agricultural Labor and Development—Facts and Theories: Comment," *Amer. Econ. Rev.*, Mar. 1967, *57*, 194–202.

J. N. Bhagwati and S. Chakravarty, "Contributions to Indian Economic Analysis: A Survey," *Amer. Econ. Rev. Supp.*, Sept. 1969, *59*, 1–73.

W. S. Comanor and H. Leibenstein, "Allocative Efficiency, X-Efficiency and the Measurement of Welfare Losses," *Economica*, Aug. 1969, *36*, 304–09.

E. Domar, "On Total Productivity and All That," *J. Polit. Econ.*, Dec. 1962, *70*, 597–608.

M. J. Farrell, "The Measurement of Productive Efficiency," *J. Roy. Statist. Soc.*, 1957, Part 3, *120*, 253–82.

M. L. Hall and C. B. Winsten, "The Ambiguous Notion of Efficiency," *Econ. J.*, Mar. 1959, *69*, 71–86.

I. Hoch, "Estimation of Production Function Parameters and Testing for Efficiency," *Econometrica*, 1955, *23*, 325–26.

———, "Simultaneous Equation Bias in the Context of the Cobb-Douglas Production Function," *Econometrica*, Oct. 1958, *26*, 556–78.

W. D. Hopper, "Allocation Efficiency in a Traditional Indian Agriculture," *J. Farm Econ.*, Aug. 1965, *47*, 611–24.

A. M. Khusro, "Returns to Scale in Indian Agriculture," *Indian J. Agr. Econ.*, Oct. 1964, *19*, 51–80.

L. J. Lau, "Applications of Profit Functions," in D. L. McFadden, ed., *The Econometric Approach to Production Theory*, Amsterdam 1971, forthcoming.

——— and P. A. Yotopoulos, "Profit, Supply and Factor Demand Functions with Application to Indian Agriculture," Memorandum No. 98, Research Center in Economic Growth, Stanford Univ., 1970.

H. Leibenstein, "Allocative Efficiency vs. 'X-Efficiency'," *Amer. Econ. Rev.*, June 1966, *56*, 392–415.

D. L. McFadden, "Cost, Revenue, and Profit Functions," in D. L. McFadden, ed., *The Econometric Approach to Production Theory*, Amsterdam 1971, forthcoming.

Y. Mundlak, "Empirical Production Function Free of Management Bias," *J. Farm Econ.*, Feb. 1961, *43*, 44–56.

——— and I. Hoch, "Consequences of Alternative Specifications in Estimation of Cobb-Douglas Production Functions," *Econometrica*, Oct. 1965, *33*, 814–28.

M. Nerlove, "Returns to Scale in Electricity Supply," in C. F. Christ et al., eds., *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, Stanford 1960, 167–98.

———, *Estimation and Identification of Cobb-Douglas Production Functions*, Chicago, 1965.

M. Paglin, " 'Surplus' Agricultural Labor and Development—Facts and Theories," *Amer. Econ. Rev.*, Sept. 1965, *55*, 815–34.

———, "Surplus Agricultural Labor and Development—Facts and Theories: Reply," *Amer. Econ. Rev.*, Mar. 1967, *57*, 202–09.

G. S. Sahota, "Efficiency of Resource Allocation in Indian Agriculture," *Amer. J. Agr. Econ.*, Aug. 1968, *50*, 584–605.

T. W. Schultz, *Transforming Traditional Agriculture*, New Haven 1964.

A. K. Sen, "Size of Holdings and Productivity," *Econ. Weekly*, Feb. 1964, *16*, 323–26.

———, "Peasants and Dualism With or Without Surplus Labor," *J. Polit. Econ.*, Oct. 1966, *74*, 425–50.

R. W. Shephard, *Cost and Production Functions*, Princeton 1953.

C. P. Timmer, "On Measuring Technical Efficiency," unpublished doctoral dissertation, Harvard Univ. 1969.

H. Uzawa, "Duality Principles in the Theory of Cost and Production," *Int. Econ. Rev.*, May 1964, *5*, 216–20.

J. Wise and P. A. Yotopoulos, "The Empirical Content of Economic Rationality: A Test for a Less Developed Economy," *J. Polit. Econ.*, Nov. 1969, *77*, 976–1004.

P. A. Yotopoulos, (1968a) *Allocative Efficiency in Economic Development: A Cross Section Analysis of Epirus Farming*, Center of Planning and Economic Research, Athens 1968.

———, (1968b) "On the Efficiency of Resource Utilization in Subsistence Agriculture," *Food Res. Inst. Stud. Agr. Econ., Trade, Devel.*, 1968, *8*, 125–35.

———, "From Stock to Flow Capital Inputs for Agricultural Production Functions: A Microanalytic Approach," *J. Farm Econ.*, May 1967, *49*, 476–91.

——— and L. J. Lau, "A Simultaneous Equation Approach to Relative Economic Efficiency," Memorandum No. 104, Research Center in Economic Growth, Stanford University 1970.

———, ———, and K. Somel, "Labor Intensity and Relative Efficiency in Indian Agriculture," *Food Res. Inst. Stud. Agr. Econ., Trade, Develop.*, 1970, *9*, 43–55.

A. Zellner, J. Kmenta and J. Drèze, "Specification and Estimation of Cobb-Douglas Production Function Models," *Econometrica*, Oct. 1966, *34*, 784–95.

The Government of India, Ministry of Food and Agriculture, *Studies in the Economics of Farm Management*, Delhi 1957–62. Reports for the year 1955–1956: Madras, Punjab, Uttar Pradesh, West Bengal; Report for the year 1956–1957: Madhya Pradesh.

Appendix

Table 1—Data for Indian Agriculture[a]

State	T	K	L	w'	Π	V
West Bengal	12.15	127.33	402.41	1.54	923.28	1,811.56
	16.96	116.00	628.37	1.61	772.36	2,403.23
	.64	7.44	39.05	1.60	187.78	129.41
	1.81	14.84	97.96	1.49	373.03	352.59
	3.11	25.19	173.10	1.59	555.87	547.05
	4.47	33.30	213.58	1.53	1,948.21	809.07
	6.18	41.59	321.42	1.45	813.20	1,158.13
	8.15	37.89	323.80	1.54	955.08	1,401.80
Madras	11.81	86.21	336.58	.54	1,653.61	907.01
	17.35	93.69	395.58	.56	2,215.54	1,174.59
	22.97	103.36	560.41	.62	2,248.45	1,683.70
	43.78	205.76	897.49	.55	5,838.73	3,607.47
	1.61	39.60	179.35	.62	426.00	354.04
	3.66	37.69	229.85	.52	716.90	751.03
	6.02	67.42	276.92	.56	2,045.88	947.55
	8.83	98.89	342.60	.56	763.14	1,190.28
Madhya Pradesh	12.44	9.57	294.70	1.08	1,709.28	1,479.12
	17.19	11.86	403.45	1.00	6,718.47	1,693.21
	24.25	14.55	470.21	1.11	40.53	2,616.57
	34.77	31.64	756.25	1.04	144.37	3,689.10
	45.17	41.10	1,084.08	1.11	157.86	4,458.28
	93.36	82.15	1,831.72	1.15	334.62	10,017.53
	2.95	3.42	101.13	.94	513.87	422.73
	7.38	8.63	190.40	1.06	729.34	849.44
Uttar Pradesh	12.00	78.00	602.40	1.06	7.57	2,448.00
	16.90	95.99	765.57	1.06	320.98	3,380.00
	27.58	148.93	1,073.14	1.01	384.68	5,653.90
	3.33	31.00	209.16	1.01	411.48	922.41
	7.68	64.97	432.84	.98	227.80	1,843.20
Punjab	14.50	19.57	450.22	1.51	448.19	2,463.55
	28.45	20.48	701.86	1.38	124.41	4,056.97
	81.19	30.85	1,484.96	1.92	391.14	12,957.92
	3.98	8.95	158.96	1.33	129.43	702.47
	7.45	7.37	270.88	1.40	377.94	1,270.22

[a] For identification of variables, see Section IV of text.

The Incidence of Social Security Payroll Taxes

By JOHN A. BRITTAIN*

One of the ironies of recent years is the sharp rise in the taxation of labor income in the midst of a declared war against poverty. As of 1970, a family of six earning a wage or salary of $5,000 per year was officially classified by government agencies as living in a condition of poverty. Its plight is taken into account by the income tax system which completely spares such a family by means of exemptions and deductions. Although the social security system refers to its own levy as a "contribution," it is in fact an involuntary payroll tax without exemptions or deductions. A tax of $240 is withheld from this $5,000 family, and families farther down in the poverty range also pay 4.8 percent. Furthermore, the additional taxes for social security (including unemployment insurance) charged against employers average about 7 percent of covered earnings, and it is important to ask who is ultimately paying this portion of payroll taxes. If, as classical theory suggests, labor really bears both parts, the total payroll tax imposed on a family in the poverty range is about 12 percent—$600 in the case of the $5,000 family. Moreover, much steeper employer payroll taxes are imposed in many other countries (e.g., about 40–50 percent in Italy), and the incidence of this type of tax is highly important in assessing labor costs to the employer and the tax burden on employees.[1]

Although there is little empirical evidence on the burden of this tax, some economists believe that it is virtually axiomatic that the employer contribution is borne by labor in the long run. For example, Milton Friedman has written:

> [The total tax for social security] includes what is euphemistically called "a contribution by the employer." Again, this is mislabeling. It is no contribution by the employer: it is a compulsory tax and it isn't paid by the employer; it is, in effect, paid by the wage earner. It is part of his wages that is sent to Washington instead of going to him. The form, the name, doesn't change the substance.[2]

Many other economists, social security specialists, employers, and labor unionists in this country appear to hold an agnostic view of the proposition that labor bears the employer tax. The tax is thus seen by some as mitigating the regressive impact of the tax on employees,[3] and the Social Security Administration has been adamant in its belief that the employer tax should not be imputed to employees in comparing

* Senior Fellow, Brookings Institution. The issue considered here arose within a general study of the economic effects of payroll taxation under a program supervised by the National Committee on Government Finance, and financed by a special grant from the Ford Foundation. I am grateful for the suggestions of Joseph A. Pechman, Joan L. Turek, the referee, the managing editor, and for the research assistance of Sheau-eung Lau and Julia Clones. The interpretation and conclusions are my own and do not necessarily reflect the views of the staff, officers, or trustees of the Brookings Institution.

[1] The concepts of incidence and shifting adopted here stress the effects of the tax on aggregate factor shares. The degree of shifting measures the extent to which the effective incidence departs from the impact incidence (see Richard Musgrave, ch. 10).

[2] See Milton Friedman, p. 8.

[3] The employee tax is generally regressive because the marginal tax rate is usually proportional below some earnings ceiling, and zero above it, and because the tax does not apply to property income which is concentrated in the higher brackets; a tax truly borne by employers would tend to offset this.

lifetime taxes and benefits under the social security system.[4] Among economists, George Jaszi has concluded: "The shifting of social security taxes is a matter about which little is actually known" (p. 53). A well-known textbook by Harold M. Groves repeats this view in a forlorn summary of the state of knowledge:

> The incidence of these taxes cannot be predicted with any great degree of assurance. On the whole, the safe conclusion would seem to be that it is divided among employers, labor and consumer; but in what proportion cannot definitely be said. [p. 157]

These conflicting opinions invite further investigation of the question, especially in view of the fact that the payroll tax has been the fastest growing major tax in the postwar United States. The maximum tax per employee for Old Age, Survivors, Disability and Hospital Insurance alone rose from $60 in 1949 to $749 in 1970 and is scheduled to reach $1,017 in 1973. Expansion of the social security tax was also the primary factor in the rise of total "Contributions for Social Insurance" from $5.7 billion in 1949 to about $54 billion in 1969. The yield from this taxation of labor income substantially exceeded corporate income taxes in 1969, and the yield was about one-half that of all individual income taxes. "After hearing for years that social security benefits are too low. . . Congressmen are now deluged with complaints that social security taxes are becoming burdensome."[5] The upward trend in payroll taxation is also spreading to the local level where city governments are showing increasing interest in payroll taxes as one means of snaring revenue from nonresidents employed in the city. Payroll taxation bulks far larger in Europe and many less industrialized countries, and Britain is now trying to move labor into manufacturing by means of a selective employment tax.

As this type of taxation grows, the unresolved question of the incidence of the employer portion (well over one-half of the *U.S.* payroll taxes) becomes increasingly important for a number of reasons. The location of the tax burden is relevant to appraisal of its effects on employment, income distribution, economic stability and growth. It is also essential to appropriate income tax treatment, appraisal of the rate of return on "contributions," rational collective bargaining, and the impact on migration and relative labor costs of competing economies.[6]

The present objective is to isolate quantitatively the impact of employer payroll taxes on factor shares in the long run. The primary approach builds on the relationship between the compensation and productivity of labor which has been frequently observed, and is also implied under certain assumptions by the specification of a production function. An intercountry analysis is undertaken which takes statistical advantage of the wide variation in tax rates among countries. The overall indication is that firms treat their payroll tax like any other labor cost in setting output and price and in agreeing to the total compensation to be awarded a given degree of labor productivity. The long-run result appears to be that employers in the aggregate avoid the burden of their contribution via a trade off between the tax and real wages and salaries.

[4] "First, it is necessary to make clear that any such comparison should not include the employer contribution, because that contribution must necessarily be considered to be pooled for the benefit of all covered persons . . ." (U.S. House of Representatives, p. 331). This reason for excluding the payroll tax is unacceptable, since the distribution of *benefits* payable out of the receipts from the employer tax is irrelevant to the question of the incidence of the tax itself. It certainly does not justify treating the tax as a burden on no one.

[5] *New York Times*, March 20, 1967, p. 15.

[6] A recent attempt to show that capital was not relatively lightly taxed in Western Europe rested entirely on an a priori assumption that payroll taxes are borne by capital. See Vito Tanzi, pp. 39–44.

I. Some Preliminary Conjecture and Casual Empiricism

Opinion and Practice in the Labor Market

Interpretation of their own behavior by participants on the economic scene is not notably reliable. However, it is worth asking how the parties in the labor market view the impact of the employer payroll tax. The issue could be expected to surface explicitly and frequently in these European countries in which the tax is much heavier than here. For example, a French union official clearly implied in a graphic opinion that the employer "contribution," in large part earmarked for family allowances, is in reality paid by labor: "We are getting paid less and less for our work, and more and more for being 'Father Rabbit,' " he complained.[7] It is clear that he was taking for granted that the tax on employers was paid at the expense of the basic wage. A similar French union opinion was reported by Lorwin (1952, p. 362): "The real incidence of social and indirect wage charges made the working class function as a vast mutual aid association in which . . . it was the poor who were helping out the poorer." And as stated in James Vadakin (p. 131), "French employers frequently argue in collective bargaining that the imposition of social charges precludes their granting of wage increases." Even in Sweden where the tax is much smaller, it appears to have forestalled wage increases; for example, in 1959:

> Swedish workers attached immense importance, during the wage talks to the pension plan which the Government was expected to adopt. . . . The unions felt, however, during the wage negotiations, that they could not get both a pension law and a substantial contractual wage increase this year, since the pension bill called for sizable employers' contributions to the pension fund, amounting to 1.89 percent of the payroll in 1960 and increasing to 4.2 percent in 1964.[8]

In the United States explicit references to a trade off between the employer tax and private compensation are less common. However, direct recognition of the shifting of the tax to employees can be found, for example, in the academic world. At 35 percent of the colleges and universities affiliated with the Teachers Insurance and Annuity Association, the institution's contribution to this private pension fund is adjusted downward to offset the cost of each increase in the taxable ceiling under the social security program. Many collective bargaining agreements concerning private pension payments have included similar offsets; however, this practice is becoming rare. Even so, since employers are likely to assess their ability to pay in terms of profits after all such costs are netted out, tax increases cut expected profits and weaken labor's case for increases in the basic wage.

Some Casual Empirical Evidence

In an informative article, Majorie W. Hald found that International Labour Office data show an inverse relationship between the basic wage and the rate of "social charges" (employer payroll tax rates) in the manufacturing sector of thirteen European countries in 1954.[9] This suggested that increased employer contributions might tend to be substituted for increases in the basic wage.[10] Similar conclusions were reached by G. R. Reid (p. 111) on the basis of data for five countries, but this sample is so small that little can be read into the numbers. Other writers have looked at labor's real income

[7] See Val Lorwin (1952) p. 362.

[8] Bureau of Labor Statistics, p. 6.

[9] See Hald, pp. 33–35. No statistical test was provided, but the Spearman rank correlation coeficient is −0.50, just significant at the 5 percent level on a one-tail test. (Only Ireland was far out of line, with a low wage level despite a low tax rate.)

[10] This is, of course, a rather crude analysis, since other factors could have produced the observed statistical association. For example, governments in low-wage countries may be under great pressure to introduce public benefit programs. The more general model to be suggested here allows for such factors.

and relative share in the national income for clues on the incidence of the employer tax. Lorwin reports the following opinion on France in the early 1950's:

> Despite understandable employer complaints about the burden of social charges, the increased benefits have come, not out of profits, but essentially out of a redistribution of income within the working class, as between direct wage and social wage recipients. . . . This phenomenon is made clearer by the fact that both (1) the total of *real* income going to wage earners (in the form of direct wages and social wage payments) and (2) wage earners' share in national income are about the same as in 1938. [1954, p. 46]

The stability of the ratio of labor's total compensation (including employer taxes) to national income, despite the great relative increase of employer taxes, is consistent with the proposition that the tax is substituted for the basic wage. However, this is a *ceteris paribus* interpretation, and other variables require consideration.

The above empirical reports are no more than suggestive. They are very persuasive however in comparison with the most recent empirical effort to appear,[11] which found that the share of property income in the national income of Puerto Rico fell with the introduction of the social security tax. This decline was deemed statistically significant on the basis of a misapplication of the chi-square procedure. This was then interpreted as showing that the employee tax (as well as the employer tax) was borne by capital even though the fall in the share of capital was four times too large to be explained by the tax alone. The reasoning and conclusions of that study are completely unconvincing.

II. Theoretical Analysis

Marginal Productivity Theory

The classical view that a universal employer tax is borne by labor was stated succinctly by Harry G. Brown (pp. 160–63). He assumes that rationality and competition lead employers to hire workers until the point is reached where the wage is just barely recouped by the marginal value product. If a tax is imposed on the employer in proportion to labor hired, the marginal worker will no longer be hired unless he accepts a reduced wage which is lower by the amount of the tax. He can be expected to do so; his labor supply function is assumed to be highly inelastic, since he cannot hold out for long and has nowhere else to go under a universal tax. Brown denies that the tax could lead to higher prices, because he assumes the tax does not increase the money supply and would have little effect on aggregate demand.[12]

Brown's underlying assumption of a fixed amount of labor supplied is subject to question. The overall effect of a tax on labor supplied cannot be forecast with any confidence, since it depends on the unknown preference functions of individual workers with respect to income and leisure. The substitution effect tends to produce a contraction of labor supplied which may be counteracted by the income effect, but the relative strength of the two forces on balance is unknown. Certainly recent empirical studies in this area have produced nothing like a consensus, and this remains a qualification of Brown's case.[13]

[11] See Elizabeth Deran. For a gentle demolition of this article see Ronald Hoffman.

[12] It is not clear why Brown rules out price increases supported by increases in aggregate money demand. If the tax proceeds are to be spent, an effort by employers to recoup the tax through price increases could be successful since the addition to government spending could yield enough total spending to buy the same output at the higher prices.

[13] Brown was considering a tax on labor with an assumed universal coverage—a structure approached fairly closely in this country. The incidence of a tax applied only to certain segments of the economy such as the British Selective Employment Tax concentrating on "services" is a more complex issue. Under certain simplifying assumptions such a tax, despite its allocative and price effects, may also be borne entirely by labor. For example, assume validity of the simple marginal productivity model, a world of two industries (one taxed, one not), two factors (labor and capital), unitary

elasticity of demand for products, Cobb-Douglas production function, fixed supplies of factors and competitive conditions; under these assumptions, it can be shown that labor's share bears the exact amount of the tax despite allocative, price, and wage rate effects.

Analysis Under Less Restrictive Assumptions

Theoretical analysis of the type outlined thus far has been criticized for over-simplified assumptions and dependence on the validity of the marginal productivity theory of wages.[14] However, it is possible to restate the argument with greater generality and with less dependence on the various assumptions of the marginal productivity theory.

The theoretical case for imputing the employer tax to labor does not depend on profit maximizing behavior by employers. The less restrictive and more plausible assumption of cost minimization is sufficient, since it requires that the ratio of marginal value product to factor cost must be the same for all factors of production. With the imposition of an employer tax in relation to labor inputs, cost minimization cannot be maintained at the given level of employment unless the tax is recouped one way or another at the expense of the basic wage.[15]

It could be argued further, however, that the shifting of the tax to labor is not even dependent on the achievement of cost minimization, or competition in the labor market. Subject to qualifications considered later, it is only necessary to accept the concept of a demand curve for labor and the idea that it makes no difference to the employer what the sums (called "total compensation" in the national accounts) which he must pay to hire labor are called. The demand curve indicates that he will pay a certain price (measured in real terms) for N units of labor of a given quality, and it makes no difference whether part of the price is called the employer's contribution and sent to Washington along with the employee's tax. There is no reason to expect a different employer reaction to the two components. Suppose the two taxes were suddenly combined into one package designated an employer tax and withheld by the employer as before. Except for short-run qualifications such as the current labor contract, there is no reason to expect the employer to pay more total compensation as the result of this accounting change. He would be required (in terms of formal accounting) to pick up the employee's part, and the nominal wage (total compensation minus employer tax) he would be willing to pay would be correspondingly reduced. A rational worker would be pleased to accept this cut in his nominal wage subject to federal and state income taxes, since the net result would be a slight increase in his after tax income. Because of the lower nominal earnings base the payroll tax and income tax paid out of the worker's given total compensation would be reduced.[16] Furthermore, this argument does not depend on competition; it is equally true in noncompetitive labor markets where labor may receive less than its marginal value product.[17]

The purpose of the above argument is to support the one basic premise that the total real compensation which can be extracted for a given amount of labor (of given quality) is independent of the labels attached to the components. If this is accepted, and in addition the aggregate labor supply curve is completely inelastic, both payroll taxes are clearly borne by labor,

[14] See Seymour Harris, pp. 291–99. See also other chapters in Part III of his work for detailed summaries and analysis of various incidence theories.

[15] The complication of the picture by any employment effects is considered later.

[16] There would be no corporate income tax effects, since wages and taxes are both deductible.

[17] Even in a labor market with monopsonistic hiring and monopolistic labor supply, where the equilibrium is indeterminate, it does not seem plausible that the employer would distinguish among the components of compensation.

and there is no effect on the cost of labor or aggregate employment. The two taxes will be paid out of the total pre-tax compensation offered to labor at the fixed level of employment. Even if the supply of labor is elastic the same result (including no aggregate substitution of capital for labor) would occur if labor bargains in terms of total compensation, including both withheld "contributions," rather than after-tax income; that is labor views both taxes as part of its earnings, just as they are definitely part of the employer's costs. In this case the tax will still have no employment effect, and therefore there is no ambiguity about the outcome.

Only if two conditions hold simultaneously does an ambiguity arise. If the supply curve for labor were not perfectly inelastic *and* the supply price excluded the employer tax, the employer's attempt to recoup the tax from the basic wage would generally produce an employment effect. This result depends on labor's viewing one withheld tax as part of its income but not the other.[18] This behavior, though difficult to rationalize, is possible, and the effect of the tax on labor's share would then depend on the elasticities of the labor supply and demand functions. Whether employment rose or fell, the basic wage bill could decline by more or less than the tax, depending on the change in the wage rate. Furthermore, we could not draw any particular conclusion such as a general tendency for labor to bear *less* than the full amount of the employer tax. This case should probably be considered a minor qualification of the general proposition, and even this minor indeterminacy depends upon an irrational view by labor that the employee tax is a price paid for future income and services while the employer tax is paid for nothing. Note that the qualification is of less significance the more inelastic the supply of labor. The a priori case that the burden of both taxes is on labor seems very strong at the aggregate level.

Some Qualifications

Two types of qualification of the above reasoning should be indicated. First, the stress on real factor shares abstracts from the actual mechanism of the shifting process and makes no distinction between "backward" (money wage effect) and "forward" (price effect) shifting. While secondary to an analysis of the distribution of real incomes, the shifting process itself may affect the competitive position of economic units and regions. For example, if employers recoup the tax by price increases, exports would decline (unless exchange rates were adjusted).

A second qualification of the present reasoning is that its aggregative focus glosses over many likely microeconomic effects of actual payroll tax systems. Varying elasticities and rate differentials may produce a highly variable impact on different economic units and allocative effects of unknown importance. Still, although no light will be shed on these microeconomic effects of the tax, the question of its overall impact on labor's share seems of sufficient importance to warrant statistical investigation.

III. Hypotheses and Regression Models

Initial Hypotheses

The degree of shifting may be defined as the fraction s of the employer tax actually borne by labor. The a priori hypothesis is that the tax is borne entirely by labor; i.e., $s=1$. This follows from a competitive model of the demand for labor and an assumed zero elasticity of aggregate labor supply with respect to the real wage rate. Under pure "backward shifting" (no pro-

[18] If the employee tax were also excluded from the supply price of labor, the same ambiguity would arise with respect to its incidence. The presumption that labor bears the employee tax only implies that the two taxes are not evaluated in the same way by labor.

duct price or employment effects from the tax), the fixed amount of labor would only be hired if the basic wage is lower by the amount of the tax than it would be in the absence of the tax. This is summarized in the first line of Table 1. If the assumption of zero supply elasticity is dropped *and* labor does not regard the employer contribution as part of its compensation, the resulting employment effect makes s dependent on the elasticities. In the special case of a unitary demand elasticity, the outcome would be determinate by definition. Since the aggregate compensation of labor (including the tax) would be invariant with respect to employment, the basic wage bill would be reduced by exactly the amount of the tax, and $s=1$ as indicated in the first column of Table 1. Thus even an employment effect does not necessarily rule out the proposition that labor bears the full burden.[19]

TABLE 1—DEPENDENCE OF HYPOTHESIZED VALUES OF THE SHIFTING COEFFICIENT s ON LABOR SUPPLY AND DEMAND ELASTICITIES

Elasticity of Labor Supply (Finite)	Elasticity of Labor Demand (Finite, Absolute Values)			
	1	>1	<1	0
0	$s=1$	$s=1$	$s=1$	—
>0	$s=1$	$s>1$	$0<s<1$	$s=0$
<0	$s=1$	$s<1$	$s>1$	$s=0$

The alternative hypotheses of $s>1$ and $s<1$ can be derived from plausible assumptions, as indicated in Table 1. Assuming a labor supply function with positive elasticity and a demand elasticity greater (less) than one, imposition of the tax would reduce the basic wage bill by more (less) than the amount of the tax.[20] The more elastic the demand, the greater the negative employment effect relative to the positive compensation rate effect. In the case of a backward sloping labor supply curve, however, these relationships are reversed, since employment would increase (ignoring the unstable case in which the supply curve is less steep than the demand curve). With a demand elasticity greater (less) than one, imposition of the tax would reduce the basic wage bill by less (more) than the tax.

While values of s other than unity can be derived from plausible elasticities, the hypothesis $s=0$ (zero wage effect of the tax) seems highly implausible. This result depends on zero elasticity of demand with respect to the wage rate (in the case of an upward sloping supply curve). As summarized in Table 1, under any demand elasticity greater than zero, imposition of the tax would cut employment and the after-tax wage bill ($s>0$). J. R. Hicks has analyzed the determinants of the elasticity of derived factor demand at the industry level in a two-factor world (pp. 241–46 and pp. 373–78). Under the assumptions of the competitive marginal productivity model there is an exact relationship between the elasticity of the derived demand for labor e_d and a the elasticity of demand for the product, b the elasticity of the supply of capital, c the relative share of labor, and d the elasticity of substitution:

$$e_d = \frac{d(a+b) + cb(a-d)}{a+b-c(a-d)}$$

A zero elasticity of labor demand is implausible in itself, and this relation re-

[19] Even this case is not entirely free of ambiguity concerning the impact of the tax, however. Assuming that firms were maximizing profits before the tax, the employment adjustment will only reduce—not eliminate the bite out of the share of capital. Thus the combined decline in the shares of labor and capital would be greater than the amount of the tax in this case of unitary demand elasticity.

[20] Since employment would be reduced by the tax, a demand elasticity greater than one would result (by definition) in a reduction in aggregate compensation (which includes the tax) and, therefore, a decline in the after-tax wage bill greater than the total employer tax ($s>1$).

inforces this a priori impression. At least two of the determinants of e_d (e.g., product demand elasticity and elasticity of substitution) must be zero in order for the expression to vanish. In view of the implausibility of such extreme values, the hypothesis $s=0$ seems untenable within the framework of the competitive model and an assumed upward sloping labor supply curve. Only if the latter is backward sloping is zero shifting possible with a demand elasticity other than zero, and this requires a special combination of elasticities. Even so, it seems in order to test $s=0$ as the logical alternative to $s=1$.

Cross-Section Regression Models

The objective here is isolation of the long-run impact of the employer tax on real wage rates, as distinct from the speed and process of shifting. Cross-section regression analysis of aggregative data for countries can offer direct evidence on the long-run response to the tax.[21] The statistical models used were variations and elaborations of the estimating equation that emerges from the constant elasticity of substitution production function (*CES*).[22] The variables originally considered in statistical analysis of the *CES* function were:

V = Value-added in thousands of *U.S.* dollars

L = Labor input in man-years

w = The basic wage rate (total labor cost excluding employer contributions for social insurance, divided by L), in dollars per man-year

The logarithmic transformation was generally favored, and the usual relationship estimated was:

$$\text{(1)} \qquad \log V/L = a_1 + b_1 \log w + u_1$$

Assuming validity of the *CES* production function, competitive product and factor markets which are in equilibrium, correct measurement of variables and an exogenous wage rate, the estimate of the slope b_1 is an estimate of the elasticity of substitution between labor and all capital inputs. The claim that the estimate of b_1 is an estimate of this elasticity has been challenged on many grounds,[23] but this interpretation of the coefficient is not essential to the present application of the model. One can acknowledge the distinguished paternity of equation (1) as a point of departure without being dependent in any way on the rigid assumptions needed to deduce it from the underlying theoretical construct. The relationship is a commonsense one a priori, and indeed the original authors apparently tried it out statistically before coming up with the *CES* function as a theoretical underpinning. In addition, the particular direction of association specified in equation (1) does not follow from the underlying model, and the reverse specification appears equally plausible. To avoid dependence on any assumption on this score, the equations have been fitted both ways.

The assumed relationship was generalized to permit explicit analysis of the impact of the employer contributions for social insurance. The new specification can be most readily rationalized for the version which states (in logarithmic form) that the average wage rate in a country is dependent on the productivity of its labor:

[21] For a statement of the argument that cross-section regressions do generally yield long-run relationships see, for example, Lawrence Klein, (1962) pp. 52–60. The stress on countries rather than states was dictated by institutional realities. There are enormous intercountry differences in employer payroll tax rates but only minute differences across the states of this country.

[22] For the original presentation of the underlying theoretical construct, see Kenneth Arrow et al, pp. 225–50.

[23] For example, it has been argued that the *CES* function itself, like the Cobb-Douglas or models with fixed input coefficients, is only a special case of a more plausible and more general model. Recognition of this would require the presence of the capital variable in the estimating equation.

(2) $\log w = a_2 + b_2 \log V/L + u_2$

The question immediately arises as to whether the wage rate associated with a given level of productivity includes the employer payroll tax per unit of labor in addition to the basic wage rate.[24] This element can be incorporated as an effective tax rate t applied to the basic (private) wage rate w. The rate of compensation of labor was assumed to be best measured by w plus some unknown fraction s of the employer payroll tax tw. This generalized model is:

(3) $\log w(1 + st) = a_3 + b_3 \log V/L + u_3$

For estimation of s, this may be rewritten with the basic wage rate as the dependent variable:

(3a) $\log w = a_3 + b_3 \log V/L - \log(1+st) + u_3$

The coefficient s in this model may be interpreted as the "shifting coefficient," or the fraction of the employer tax per worker which is actually borne by labor. For example, if s should equal zero, $\log(1+st)$ equals zero, and the estimated basic wage rate would be independent of the tax, depending solely on productivity; this would indicate no shifting. However, an s value of unity would indicate that for a given level of productivity V/L the presence of the tax lowers the estimated basic wage rate by just the amount of the tax.[25] This shows a direct and complete trade off between the basic wage rate and the tax per worker, or a 100 percent shifting of the tax burden at the expense of labor's basic wage.[26]

The term $\log(1+st)$ is awkward for estimation purposes, but the parameter s can be extricated by approximating $\log(1+st)$ either by $s \log(1+t)$, or by the first term of the Taylor expansion $s(.434t)$;[27] results by the two methods agree closely. For estimation purposes two approximations of the last equation emerge:

(4) $\log w = a_4 + b_4 \log V/L - s_4 \log(1+t) + u_4$

(5) $\log w = a_5 + b_5 \log V/L - s_5 (.434t) + u_5$

Treatment of $\log V/L$ as the dependent variable leads by analogous reasoning to two alternative estimating equations in which s continues to be interpreted as the shifting coefficient:

(6) $\log V/L = a_6 + b_6 \log w + b_6 s_6 \log(1 + t) + u_6$

(7) $\log V/L = a_7 + b_7 \log w + b_7 s_7 (.434t) + u_7$

These four models all rely on the logarithmic transformation of the variables. This form is preferred and stressed for methodological reasons.[28] However, the corresponding models without the transformation were also estimated in the early

[24] Note that the basic wage rate w, as defined above, includes private fringe benefits and does not exclude payroll taxes paid by employees.

[25] If $\log w_0$ and $\log w_1$ are the regression estimates of the dependent variable before and after tax (and w_0 and w_1 are the implied basic wage rates), $\log w_1 = \log w_0 - \log(1+t)$, and $w_0 = w_1 + tw_1$. So, the new basic wage rate tends to equal the original basic wage rate less the tax per unit of labor. This interpretation may be generalized to cover values of s other than zero or one. The model implies that $w_1 = w_0 - stw_1$; the new wage rate tends to fall short of the old by the fraction s of the tax tw_1.

[26] This shifting interpretation is, of course, subject to the qualification concerning possible employment effects discussed in Section II. The effect of the tax on the wage *bill* could be different in relative terms from its effect on the wage rate. However, any such complication signals no particular direction of bias in s as a measure of the degree of shifting.

[27] The first of these approximations is exactly correct when s equals zero or unity, too low for s values in between, and too high for s greater than one. The second is correct if $s=0$ and too high for positive values of s. The absolute error in the approximations varies positively with t but does not seem excessively large for extreme values of t such as 0.4. In any case the results from both approximations will be presented to suggest the order of magnitude of error.

[28] From the theoretical point of view, the logarithmic version might be favored because it is the one which emerges from the well-known *CES* rationale. It was also preferable for statistical reasons, since it afforded a closer approach to the property of homoscedasticity.

stages, because they avoided the need to approximate $log(1+st)$.[29] The equations are

$$(8) \quad w = a_8 + b_8 V/L - s_8 tw + u_8$$

$$(9) \quad V/L = a_9 + b_9 w + b_9 s_9 tw + u_9$$

Models (4)–(9) offer six alternative estimates of the shifting coefficient s.

IV. Empirical Findings

Most of the basic country data on value-added, wages and salaries, and employment were taken from the manufacturing censuses for 1958, but any annual census in the period 1957–59 was treated as eligible for the cross-section analysis. Three sets of countries and four sets of currency conversion ratios (called x_1, x_2, x_3, and x_4) were used in fitting the models. Effective payroll tax rates were estimated from statutory rates. Data sources, criteria for country selection and methods of calculating the conversion ratios are outlined in the Appendix.

Table 2 reports estimates of the shifting coefficient s based on models (4)–(9) fitted to data for aggregate manufacturing.[30] All of the models produce close fits for all sets of data, with 92–96 percent of the variance explained in each case.[31] The results for the logarithmic models (4)–(7) show positive values of s which are significantly greater than zero in every case (at the 2.5 percent level of significance or better). On the basis of this methodology, the no-shift hypothesis appears thoroughly discredited at the aggregate level. Although all 24 of the point estimates of s (falling in the range 1.14–1.60) are greater than the alternative hypothetical value of unity, they are not embarrassingly large; the estimates exceed one by a maximum of about one standard error and are therefore consistent with the hypothesis that 100 percent of the tax is borne by labor. The results for models (8) and (9) without the logarithmic transformation are somewhat more erratic, but tell the same story. These models show estimates of s significantly greater than zero in ten out of twelve cases. In only two cases were the estimates substantially greater than unity. In sum, models (8)–(9) support the shifting results of models (4)–(7) but less convincingly.

Wages and productivity in two-digit manufacturing industries were next analyzed for the maximum number of countries with data available; results for individual industries are presented in Table 3.[32] Several of the industry models fit considerably less well than the aggregate models, but on the whole the $\bar{R}^2$ estimates remain very high (falling below 0.9 for only three industries). The estimates of the shifting coefficient s based on models (4)–(7) are significantly greater than zero at the 5 percent level or better in the case of 7 of the 13 industries, but in no case significantly greater than unity; thus the hypothesis that $s=0$ can be rejected in favor of $s=1$ is supported strongly in the majority of the industries. Only in the single case of the nonmetallic mineral products industry can the hypothesis that $s=1$ be rejected in

[29] Equation (8) also offers the most direct interpretation of the coefficient s. For example, if the estimate of s turns out to be unity, the expression says that for a given level of productivity the higher the tax per unit of labor tw, the lower the estimated basic wage rate by the same amount.

[30] The fitting of models to aggregative data is clearly a distasteful procedure when viewed within the production function framework. However, the negative association revealed between the tax variable and the basic wage, given productivity, nevertheless seems meaningful and indicative of a trade off between basic wages and employer taxes.

[31] Results are reported for conversion ratios x_1 and x_2 only, because they were very nearly duplicated by results for ratios x_4 and x_3, respectively. It is also worth noting that the models using estimated purchasing power parity ratios x_1 and x_4 produced closer fits in every case than those for the relatively arbitrary official exchange rates contained in variables x_2 and x_3.

[32] The purchasing power parity ratio x_1 produced closer fits than the exchange rate x_2 for all 13 industries. However, the estimates for s were so similar in the two cases that only those for x_1 are tabulated.

Table 2—Regression Estimates of Shifting Coefficients Based on Aggregate Data for All Manufacturing and Conversion Ratios x_1 and x_2

Model	Statistic	64 Countries (Sets A, B, C)		44 Countries (Sets A, B)		30 Countries (Set A)	
		x_1	x_2	x_1	x_2	x_1	x_2
(4)	s	1.325[c]	1.326	1.535[c]	1.435[b]	1.538[b]	1.564[b]
	$S(s)$	(.463)	(.467)	(.463)	(.486)	(.650)	(.691)
	$\bar{R}^2$	.930	.924	.934	.921	.944	.926
(5)	s	1.149[c]	1.140[c]	1.317[c]	1.224[c]	1.392[b]	1.413[b]
	$S(s)$	(.408)	(.411)	(.405)	(.425)	(.580)	(.616)
	$\bar{R}^2$	.929	.924	.933	.920	.944	.926
(6)*	s	1.517[c]	1.533[c]	1.597[c]	1.471[c]	1.527[b]	1.561[b]
	$S(s)$	(.473)	(.478)	(.478)	(.507)	(.673)	(.721)
	$\bar{R}^2$	.932	.926	.934	.920	.934	.925
(7)*	s	1.313[c]	1.313[c]	1.370[c]	1.252[c]	1.380[b]	1.407[b]
	$S(s)$	(.416)	(.421)	(.418)	(.444)	(.600)	(.644)
	$\bar{R}^2$	.931	.926	.933	.920	.943	.925
(8)	s	1.286[c]	.892[a]	1.484[c]	1.004[a]	1.046[a]	.658
	$S(s)$	(.395)	(.454)	(.424)	(.530)	(.556)	(.678)
	$\bar{R}^2$	.943	.932	.948	.947	.960	.934
(9)*	s	1.706[c]	1.428[c]	1.762[c]	1.418[b]	1.310[b]	1.083
	$S(s)$	(.386)	(.452)	(.414)	(.532)	(.550)	(.685)
	$\bar{R}^2$	.949	.938	.953	.933	.963	.937

Source of data: See Appendix.

* The coefficient s in models (6), (7), and (9) is estimated by the ratio of the coefficients bs and b. The standard errors $S(s)$ for these three models are derived from an abbreviated version of the Taylor's expansion approximation (see Klein (1953), p. 258). The estimate used was the ratio $S(bs)/b$. This approximation ignored the variance of b and the covariance of b and bs on the ground that they were generally small relative to the variance of bs.

[a] Different from zero in the expected (positive) direction at the 5 percent level of significance.

[b] Different from zero in the expected (positive) direction at the 2.5 percent level of significance.

[c] Different from zero in the expected (positive) direction at the 0.5 percent level of significance.

favor of the hypothesis that $s=0$. The median estimate of s is that obtained for the clothing and footwear industry, whichever one of the four models is applied; though each of these estimates of the shifting coefficient s is slightly greater than one, none is significantly above. These industry results, while not unanimous, greatly favor the hypothesis $s=1$ over $s=0$ and indicate a 100 percent trade off between payroll taxes and the basic wage rate.

In an attempt to pin down the estimate for the aggregate shifting coefficient for manufacturing, the industry data underlying Table 3 were pooled for the final estimates of models (4) and (5). Dummy variables were introduced to permit the constant term and coefficient of $log\ V/L$ to vary by industry.[38] Results for the two models and two exchange conversion methods are presented in Table 4.

The dummy variable technique allows 380 degrees of freedom and should yield by far the most accurate estimates of s attained in this study, with standard errors reduced to about 0.25. The estimated values of the shifting coefficient in Table 4 are greater than zero by 4.0–4.5 standard

[38] Models (6) and (7) could not be treated in this way to yield a unique estimate of s because the estimate of s depended on the ratio of the single estimate of bs to the estimates of b which were assumed to vary by industry.

TABLE 3—ESTIMATES OF THE SHIFTING COEFFICIENT s BASED ON INDUSTRY DATA AND CONVERSION RATIO x_1

Industry	Number of Countries	Statistics	Model (4)	Model (5)	Model (6)	Model (7)
Food, Beverages and Tobacco	36	s	1.74[a]	1.57[a]	1.83[a]	1.65[a]
		$S(s)$	(.86)	(.78)	(.90)	(.81)
		$\bar{R}^2$	.907	.907	.907	.907
Textiles	34	s	1.82[b]	1.64[b]	1.99[b]	1.78[b]
		$S(s)$	(.86)	(.77)	(.89)	(.79)
		$\bar{R}^2$	.929	.929	.930	.930
Clothing, Footwear, etc.	31	s	1.23[b]	1.12[b]	1.23[b]	1.12[b]
		$S(s)$	(.52)	(.46)	(.53)	(.47)
		$\bar{R}^2$	.971	.972	.971	.971
Wood and Products	36	s	1.64[b]	1.52[b]	1.82[b]	1.67[b]
		$S(s)$	(.79)	(.71)	(.81)	(.72)
		$\bar{R}^2$	.933	.933	.934	.934
Pulp and Paper Products	28	s	.77	.71	.73	.67
		$S(s)$	(1.00)	(.89)	(1.04)	(.93)
		$\bar{R}^2$	.918	.918	.917	.918
Printing and Publications	34	s	1.37[a]	1.20[a]	1.43[b]	1.25[a]
		$S(s)$	(.67)	(.60)	(.69)	(.61)
		$\bar{R}^2$	.946	.946	.946	.946
Leather and Products	25	s	−.16	−.08	−.34	−.23
		$S(s)$	(.79)	(.70)	(.82)	(.72)
		$\bar{R}^2$	.929	.929	.929	.929
Rubber Products	27	s	1.15	1.07	1.13	1.05
		$S(s)$	(1.36)	(1.20)	(1.46)	(1.29)
		$\bar{R}^2$	.859	.859	.858	.859
Chemicals and Products	31	s	.89	.80	1.16	1.01)
		$S(s)$	(1.60)	(1.43)	(1.80)	(1.61
		$\bar{R}^2$	.771	.771	.772	.772
Nonmetallic Mineral Products	37	s	−.33	−.24	−.32	−.23)
		$S(s)$	(.71)	(.64)	(.73)	(.66)
		$\bar{R}^2$	.942	.942	.942	.942
Basic Metals	25	s	2.07[a]	1.86[a]	2.20[a]	1.97[a]
		$S(s)$	(1.14)	(1.02)	(1.22)	(1.08)
		$\bar{R}^2$	.879	.879	.879	.879
Metal Products	35	s	1.56[b]	1.40[b]	1.67[b]	1.48[b]
		$S(s)$	(.69)	(.62)	(.70)	(.63)
		$\bar{R}^2$	.943	.943	.944	.944
Other Manufacturing	28	s	1.02	.92	1.04	.93
		$S(s)$	(1.17)	(1.04)	(1.22)	(1.09)
		$\bar{R}^2$	.907	.908	.907	.907

Source of data: See Appendix.
[abc] For levels of significance, see Table 2.

rrors, permitting the hypothesis that $s=0$ o be rejected at the 0.003 percent level or etter. Again, although each estimate falls lightly above unity, the excess is far from ignificant, and the results strongly support the hypothesis that in the aggregate he entire employer tax is shifted to abor.[34,35]

V. Some Implications of the Findings

It should be reiterated at this point that this analysis has shed no light on the mech-

[34] It is possible that the s coefficients contain a slight upward bias for another reason. Insofar as other indirect taxes, such as the sales tax, bear more heavily on labor income than on overall value-added at factor cost *and* such taxes are correlated with payroll taxes across countries, the regressions on the payroll tax variable alone could pick up some of the influence of these other taxes; this would impart an upward bias to the estimates of s. However, the payroll tax is undoubtedly the domi-

TABLE 4—ESTIMATES OF THE SHIFTING COEFFICIENT s FOR ALL MANUFACTURING ON POOLED INDUSTRY DATA, WITH DUMMY VARIABLES FOR INDUSTRIES

(407 Observations and 380 Degrees of Freedom)

Conversion Ratio	Statistic	Model (4) Plus Dummies	Model (5) Plus Dummies
x_1	s	1.144	1.043
	$S(s)$	(.263)	(.235)
	$\bar{R}^2$	.911	.911
x_2	s	1.176	1.070
	$S(s)$	(.268)	(.239)
	$\bar{R}^2$	.904	.904

Source of data: See Appendix.

anism through which the real burden of the tax on employers is shifted to employees. Earnings and productivity variables for each country were measured in dollars by several conversion methods. The essence of the finding here is that given the level of productivity in a country, the presence of a payroll tax on employers tends to reduce the wage in dollars by roughly the amount of the tax. This could be due to a lag in the basic wage (measured in local currency) in response to imposition of the tax (backward shifting); it could be due to price increases reducing the real value of wages (forward shifting). More likely, the outcome is achieved through a combination of the alternative employer reactions. The nature of the blend is of little significance for a study of income distribution, but the extent of price adjustment does affect international competitive positions until offset by exchange rate adjustment. In any case, whichever shifting mechanism dominates, the real burden of the tax falls on labor; this has important implications however it comes about.

If the conclusion that both employer and employee payroll taxes are borne by labor is accepted, several significant corollary propositions follow. In the first place, assuming no aggregate employment effects, this suggests that payroll taxes are neutral with respect to the allocation of capital and labor in the aggregate and within a given industry.[36] If the tax has no net cost impact for employers, it produces no incentive to substitute capital for labor. The conclusion that its burden falls on labor shields the tax from the usual criticism that it promotes automation and aggravates the unemployment problem.

The economic stabilization properties of the tax are also affected by its incidence. Although it is generally conceded that the typical payroll tax is a relatively weak stabilizer, it would probably be even weaker if borne by capital. For example, assuming a higher marginal propensity to spend income from labor than income from capital, a fall in the tax on labor in response to a wage decline would be a more effective brake to limit the decline in spending than an equal decline in a tax on profits. A tax borne by labor would also presumably produce a lesser drag on growth than a tax on profits, which would cause a greater cut in saving.

Acceptance of the labor burden hypothesis is also relevant in the collective bargaining arena. It should be recognized on both sides that the employer payroll tax is just as clearly a component of the cost of hiring labor as private fringe benefits or the nominal wage itself. Labor would then regard the employer contribution as part of its compensation which is being paid in lieu of a higher nominal wage. Recognition of a trade off between wages and fringes on the one hand and employer contribu-

nant tax concentrating on labor, and the impact of other taxes should be relatively small.

[35] Various forms of regression models allowing for lagged response were applied to *U.S.* time-series data. The findings gave considerable support to the cross-country conclusions but were not conclusive enough to merit reporting here.

[36] A payroll tax which varies across industries (such as the British Selective Employment Tax) can be expected to affect the allocation of labor among industries without reducing the overall capital-labor ratio.

tions on the other would bring into more explicit focus the pros and cons of fueling social programs by this type of tax.

The appropriate treatment of labor income under the personal income tax also depends on payroll tax incidence. At present the employee pays tax on the income from which employee contributions are withheld but not on the income from which the employer tax is withheld. If the latter income is part of labor's share it should be taxed just as the source of the employee tax is; or both parts should be exempt in favor of a tax on benefits which are now exempt.

If labor ultimately pays the employer tax, this is also highly relevant to relative international competitive positions. Countries such as Italy and France with large "social charges" are not placed at a competitive disadvantage vis-à-vis countries where the employer tax constitutes only a small part of total compensation.[37] This has significant implications for tariff policy and for attempts to improve international economic cooperation.

On the previous counts the labor-burden finding implies no significant critique of the payroll tax. However, the incidence of the tax is highly significant for evaluation of its effect on income distribution. The conclusion that labor bears the tax makes clear that its burden on low income groups is greater than generally realized. It also implies that its impact on income distribution is typically regressive. These qualities of the payroll tax offer a solid basis for proposing that this form of taxation be curtailed or eliminated. This could be done by introducing exemptions of low incomes as under the income tax or substituting the income tax for all or part of the payroll tax.

Finally the incidence of the tax is significant for evaluation of the terms of social security programs. The finding that labor bears the tax points to a lower rate of return on contributions to participants in social security than if the employer tax could be ignored. It is difficult to understand the position of the Social Security Administration which has conceded that this tax is largely borne by labor in the aggregate and yet ignores it in evaluating the tax paid by individuals. It does so on the ground that no exact imputation of the tax is possible. However, if it is paid by employees as a group, it must also be paid by them as individuals, and it seems better to make imperfect imputations which are roughly right than to settle for being precisely wrong. The implication of these imputations is that wage and salary earners pay the entire tax for Unemployment Insurance and twice as much under the OASDHI program as the amount withheld from their nominal earnings. An awareness of this on the part of taxpayers might contribute to decreased reliance on this regressive form of taxation.

Appendix

Virtually all of the census of manufactures data on wages and salaries, value-added, and employment were taken from the United Nations publication, *The Growth of World Industry, 1938–61.* For a few countries, information was taken from other United Nations publications, *The Statistical Yearbook*, and *The Yearbook of National Account Statistics.*

The effective employer tax rate *t* was estimated from statutory rates in the Social Security publication *Social Security Programs Throughout the World.* Five types of employer payroll taxes were included; they were ear-marked for: (1) old-age, invalidity and survivors insurance and related programs, (2) health and maternity insurance,

[37] Even if the shifting of the burden to labor were accomplished primarily via price increases, this should have only temporary effects. The essential fact remains that the tax does not increase the real cost of labor; exchange rate adjustments could restore the competitive position which existed before any forward shifting.

(3) unemployment insurance, (4) family allowance programs, and (5) work injuries insurance. Estimates of effective rates took account of the taxable ceiling in each country. The statutory rate was adjusted on the basis of a graphical relationship between percentage of earnings taxable and the ratio of the ceiling to mean earnings, as observed in the United States. In the case of those countries that also specified minimum taxable income each earner was assumed to earn at least the minimum, and the effective rate was adjusted downward on the basis of the fraction of the total wage bill that was exempt by the minimum.

Four alternative sets of conversion ratios were used—x_1, x_2, x_3, and x_4—to convert currencies into dollars. The estimates x_1 and x_4 were "purchasing power parity ratios" based on price indexes; they were estimated by the United Nations, see *The Growth of World Industry, National Tables*, pp. 310–11, and *International Analysis and Tables*, Table 9B. These conversion ratios produced generally closer fits than the more arbitrary sets of official exchange rates x_2 and x_3. These two sets (which differ slightly due to alternative treatment of multiple exchange rates) were based on U.S. Department of Commerce data and the U.N. *Statistical Yearbook*, Table 9A.

Countries were selected for analysis if data on value-added V, wages and salaries W, and employment L were available for aggregate manufacturing in at least one of the years 1957–59; most of the censuses were for the year 1958. A few additional countries were included for which the census fell just outside the 1957–59 period and for which the "number engaged" was available rather than the number of employees. This yielded data on aggregate manufacturing for sixty-four countries labeled sets A, B, C in the text. After fitting the models to these 64 observations, the 20 countries with the smallest total wage bills, set C, were dropped and the models refitted. Finally countries in set B with data on number engaged only, with surveys outside the 1957–59 period or without data available in the main source were dropped, leaving 30 observations. This process was an attempt to utilize successively more reliable data while sacrificing observations. However, results for the three different sets of aggregative data all gave results consistent with the hypothesis that the overall shifting coefficient equals unity, as shown in Table 2.

The industry analysis in Tables 3 and 4 covers all countries among the original sixty-four for which data were available on an industry basis.

REFERENCES

K. Arrow, H. B. Chenery, B. S. Minhas, R. M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Statist.*, Aug. 1961, *43*, 225–50.

H. G. Brown, *The Economics of Taxation*, New York 1924.

E. Deran, "Changes in Factor Income Shares Under the Social Security Tax," *Rev. Econ. Statist.*, Nov. 1967, *49*, 627–30.

M. Friedman, "Transfer Payments and the Social Security System," *The Conference Board Record*, New York, Sept. 1965, *11*, 7–10.

H. M. Groves, *Financing Government*, 6th ed., New York 1965.

M. W. Hald, "Social Charges in the EEC Countries: Some Economic Aspects," *Econ. Int.*, Nov. 1959, *12*, 677–96.

S. Harris, *Economics of Social Security*, New York 1941.

J. R. Hicks, *Theory of Wages*, 2nd ed., New York 1963.

R. F. Hoffman, "Factor Shares and the Payroll Tax: A Comment," *Rev. Econ. Statist.*, Nov. 1968, *50*, 506–08.

G. Jaszi, "A Critique of the United States Income and Product Accounts," Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 22, Princeton 1958, p. 402.

L. Klein, *A Textbook of Econometrics*, New York 1953.

———, *Introduction to Econometrics*, New York 1962.

R. Lester, *Economics of Unemployment Compensation*, Princeton 1962.

V. R. Lorwin, "France: History of Trade Union Developments," in W. Galenson, ed., *Comparative Labor Movements*, New York 1952, 313–409.

——— *The French Labor Movement*, Cambridge 1954.
R. Musgrave, *The Theory of Public Finance*, New York 1959.
G. R. Reid, "Supplementary Labour Costs in Europe and Britain," in G. R. Reid and D. J. Robertson, eds., *Fringe Benefits, Labour Costs and Social Security*, London 1964.
V. Tanzi, "Tax Systems and Balance of Payments: An Alternative Analysis," *Nat. Tax. J.*, Mar. 1967, *20*, 39–44.
J. C. Vadakin, *Family Allowances*, Miami 1958.
International Labour Office, *The Cost of Social Security, 1958/1960*, Geneva 1964.
New York Times, Mar. 20, 1967.
United Nations, *Growth of World Industry, 1938–61, International Analysis and Tables*, New York 1963.
———, *Statistical Yearbook 1965*, New York 1967.
——— *The Growth of World Industry 1938–61, National Tables*, New York 1963.
———, *Yearbook of National Account Statistics, 1963*, New York 1965.
U. S. Bureau of Labor Statistics, *Labor Development Abroad*, Washington, August 1959.
U.S. Department of Commerce, *Statistical Abstract of the United States*, Washington 1958, 1959, and 1960.
U.S. Department of Health, Education, and Welfare, *Social Security Programs Throughout the World, 1958*, Washington.
U.S. House of Representatives, Hearings before the Committee on Ways and Means, *President's Proposals for Revision in the Social Security System*, Mar. 1, 2, and 3, 1967, Washington.

Determinants of the Commodity Structure of U.S. Trade

By ROBERT E. BALDWIN*

Nearly twenty years ago Wassily Leontief made the surprising discovery that a lower capital-labor ratio was required to produce a representative bundle of *U.S.* exports than was involved in producing a representative bundle of import-competing goods. The Leontief results and those from similar investigations of other countries[1] effectively destroyed the comfortable confidence of economists in the simple version of the Heckscher-Ohlin trade theory that had long been accepted mainly on the basis of casual empiricism.[2] However, the "Leontief paradox" also stimulated extensive theoretical and empirical research directed at providing alternative explanations for the commodity-pattern of a country's trade. The purpose of this paper is to test the main alternative hypotheses that have been advanced for this purpose, as well as the simple Heckscher-Ohlin theory itself, by using 1958 *U.S.* labor, capital, and input-output coefficients rather than the 1947 coefficients employed by Leontief. Information from the *1/1000 Sample of the Population of the United States, 1960* plus certain other data relating to the quality of the labor force and various structural characteristics of *U.S.* industries are also utilized in testing these hypotheses. In addition, the *U.S.* trade pattern for 1962 rather than 1947 or 1951 is used in making the various calculations.

The Heckscher-Ohlin theorem states that a country's exports use intensively the country's relatively abundant factors. As is well known (see the survey article by Jagdish Bhagwati 1965), a set of sufficient conditions for the theorem are: 1) identical production functions throughout the world for each commodity as well as qualitatively identical productive factors; 2) production functions homogeneous of degree one with diminishing marginal productivity for each factor; 3) nonreversibility of factor intensities; 4) identity of consumption patterns (in the sense that all goods are consumed in the same proportions) among countries at any given set of international commodity prices; and 5) perfect markets, free trade, no transport costs, and complete international immobility of productive factors. If one adds the condition that there are at least as many commodities as factors and that all countries produce some of each commodity, it also follows that factor-price equalization is achieved.[3] In addition, as

* Professor of economics, University of Wisconsin, Madison. Research for the paper was financed by a grant from the National Science Foundation.

[1] See Masahiro Tatemoto and Shinichi Ichimura, Donald F. Wahl, Ranganath Bharadwaj, and Karl W. Roskamp.

[2] The simple Heckscher-Ohlin theory is a model in which only capital, labor, and natural resources are the factors of production, and in which such factors as economies of scale and differences in technology do not play a part in determining comparative-cost differences among nations. It should be recognized, however, that Bertil Ohlin, even when presenting what he regarded as a simplified version of his model, divided labor into three skill groups and capital into long-term and short-term capital. Nevertheless, over time most economists have come to label trade models with a two- or three-factor breakdown as simplified versions of the Heckscher-Ohlin theory. More important, they have believed that the broad pattern of a country's trade could be adequately explained with a simple two- or three-factor model. See, for example, Karl-Erik Hansson.

[3] Factor-price equalization can, of course, be achieved without the identical-taste assumption. Moreover, it is

Pranab Bardhan has pointed out, the Heckscher-Ohlin theorem holds as long as one country's production functions all differ from those used in the rest of the world by only a neutral efficiency factor.

Obviously, the various conditions required for the Heckscher-Ohlin proposition to be logically valid do not all hold in the real world. However, this does not necessarily mean that the Heckscher-Ohlin theory is a poor theory. If the failure of any of the assumptions to hold does not systematically and significantly bias the conclusions of the model, the theory will generally still accurately predict the nature of trade patterns from a knowledge of relative factor endowments and thus be a "good" theory. Tests such as the one undertaken by Leontief are designed to determine the adequacy of the theory in this sense.

I. Alternative Explanations of the Leontief Paradox

Fortunately, trade theory has not suffered from a lack of suggested explanations for the Leontief results. Instead, the problem has been to discriminate among the several hypotheses that have been advanced to account for them. Six major (not necessarily mutually exclusive) groups of explanations can be distinguished.[4] These maintain that the actual structure of *U.S.* trade can be accounted for mainly by: 1) the relative abundance of skilled labor in the United States; 2) an efficiency advantage in favor of the United States in Research and Development (R and D) oriented industries; 3) the scarcity of natural resources in the United States coupled with a complementary relationship between natural resources and capital; 4) factor-intensity reversals sufficiently extensive to upset the Heckscher-Ohlin proposition; 5) a strong *U.S.* demand bias in favor of capital-intensive goods so that these are imported even though the United States is capital-abundant; 6) high tariffs and other trade-distorting measures that favor the domestic production of labor-intensive products and consequently bias the import bundle against these products.

Skilled Labor

Current interest in the general topic of investment in human resources has served to focus considerable attention recently (see articles by Peter Kenen (1968), Bharadwaj and Bhagwati, and Roskamp and Gordon McMeekin), on the first explanation mentioned above. This explanation was initially put forth by Leontief and Irving Kravis, both of whom pointed out that *U.S.* export industries employed more highly skilled labor than did import-competing industries. Donald Keesing (1965, 1966, 1968), Kenen (1965), Helen Waehrer, and Merle Yahr have since elaborated both analytically and empirically upon the significance of differential supplies of labor-skills and also demonstrated the importance of this factor for explaining trade patterns of other countries. Kenen has performed the interesting experiment on *U.S.* data of capitalizing the excess of wages earned by various types of skilled labor above the wages of unskilled laborers in order to obtain an estimate of the value of human capital involved in export- and import-competing production. When the estimates of human capital obtained by discounting at less than 12.7 percent are added to Leontief's physical capital estimates, the paradox is reversed.[5]

A drawback of computing human capi-

not necessary for factor-price equalization that each country produce some of every commodity.

[4] See Gary C. Hufbauer for a classification that further refines some of the categories listed.

[5] However, similar estimates by Bharadwaj and Bhagwati for India have the effect of operating against what would be predicted by the Heckscher-Ohlin theory.

tal by capitalizing income differentials at a single discount rate is, as Kenen notes, that the method assumes all income differences to be the result of differences in education and other forms of human investment. It also assumes that long-run equilibrium conditions prevail in capital markets.[6] There is considerable evidence that market imperfections due to various economic and social factors as well as differences in ability are significant explanatory variables for earning differentials. Moreover, returns to low levels of education are considerably greater than to high levels of education.[7]

An even more important point is whether it is proper to combine estimates of human and physical capital to determine the capital-labor ratio in trade-oriented production. Such a procedure rests upon the assumption that in the long run, capital moves freely between physical goods and human agents of production. This assumption may be acceptable for a highly developed country like the United States, but it does not seem appropriate for most developing nations where market imperfections even make it difficult to regard all physical capital as fungible in the long run.

R and D Oriented Industries

Keesing (1968) together with Raymond Vernon, William Gruber, and others, has pointed to the significance of research activities in explaining trade patterns. In particular, they found that there is a strong positive correlation between the relative importance of R&D activities in American industries and the exports of American industries as a proportion of total exports of all the major trading countries. These results confirmed the hypothesis that R&D expenditures are a proxy for temporary, comparative-cost advantages provided by the development of new products and productive methods.[8] A more direct method of introducing non-uniform, efficiency differences between domestic and foreign production functions has been followed by Gary Bickel in a study of *U.S.*-Japanese trade. Bickel used differences among *U.S.* and Japanese industries in the productive efficiency of capital and labor (derived from empirically estimated *CES* production functions) as an explanatory variable for differences between the two countries in relative commodity prices. He found that at least 25 percent of the total variation in these prices was attributable to the efficiency factor alone.[9]

Scarcity of U.S. Natural Resources

The third-factor (natural resources) explanation of Leontief's results has been put forth both by Muhammad Diab and by Jaroslav Vanek (1963). Vanek accepts the notion that the United States is capital abundant but states that natural resources and capital are complementary. Therefore, since natural resources are scarce in the

[6] By excluding the human capital in laborers, Kenen's method understates the total human capital involved in export- and import-competing production. Since laborers are more important in import-competing than export production, this exclusion has the effect of tending to reverse the paradox.

[7] See Gary Becker (pp. 124–27) and Giora Hanoch.

[8] This hypothesis also assumes that current R&D expenditures are representative of the stock of innovations that are the source of comparative-cost advantages and that the rate at which innovations are copied is approximately the same among industries.

[9] Leontief's explanation (1953, p. 344) of his findings, namely that *U.S.* labor is some three times more efficient than foreign labor, implies that the efficiency advantage of the United States is highly biased towards saving labor whereas the usual intercountry studies that estimate the elasticity of substitution explicitly assume only factor-neutral efficiency differences. Bickel, for example, uses the neutral efficiency parameters calculated by Kenneth Arrow, Hollis Chenery, Bagicha Minhas, and Robert Solow. There is evidence suggesting that technical progress in the United States has in fact been labor-saving. See Paul David and Th. van de Klundert. However, the very large advantage in favor of *U.S.* labor suggested by Leontief does not seem to be supported by direct studies of comparative labor efficiency. See the article by Mordechai Kreinin.

United States, both capital and natural resources are conserved through trade. This hypothesis seemed to receive support from calculations that Leontief made in his second article (1956) on the subject. Specifically, when nineteen natural resource industries were excluded from the matrix, the paradox was eliminated.[10]

As William Travis has pointed out (pp. 94–99), this explanation is logically inconsistent with a Heckscher-Ohlin model in which factor-price equalization is achieved.[11] Even though there are other general or specific factors besides labor and capital, a country that is capital abundant relative to the rest of the world will export capital-intensive products compared to its import-competing production. This can be seen in the following way. In a factor-price equalization model where all goods are traded and tastes are identical, the factor proportions used to produce any particular commodity are the same in all countries, and each country consumes all commodities in the same proportions. This implies that each country indirectly consumes each factor in the same proportions. In other words, one can think of each country as starting with given factor supplies and then trading these at common factor prices until a common set of factor-consumption ratios is reached. In the case where natural resource industries are capital-intensive and natural resources are scarce in a capital-abundant country, factor equilibrium may be achieved in two ways: by the capital-abundant country exporting items that are even more capital-using than natural resource products; or by the rest of the world, which is labor-abundant, exporting highly labor-intensive commodities in addition to the capital-intensive, natural resource products.[12]

Reversals of Factor Intensity

One of the most potentially damaging arguments against the Heckscher-Ohlin theory is that factor-intensity crossovers are extensive within relevant ranges of factor prices. Under these circumstances a country's exports to the rest of the world and the rest of the world's exports to the country may be either both capital-intensive or both labor-intensive. Then the Heckscher-Ohlin relationship cannot possibly hold for both trading units. A study by Minhas seemed to indicate that factor-intensity crossovers were in fact extensive. However, subsequent analysis by Leontief (1964), using additional data provided by Minhas, found extremely little evidence of factor-reversals. Several other recent studies[13] also failed to support the Minhas position, but the matter cannot as yet be regarded as finally settled.[14] For example, see the issues raised by Michael Hodd.

[10] The industries excluded by Leontief tend to be those in which the direct and indirect factor content of immobile natural resources is relatively high. His exclusion of all agricultural industries except livestock and livestock products seems questionable under this criterion.

[11] Vanek (1968) also has now proved this proposition rigorously. Travis (p. 97) further notes that the Leontief results are not consistent with the natural resource explanation even in the absence of factor-price equalization.

[12] Suppose that country *A* possesses a relatively abundant supply of capital, *K*, compared to country *B* (the rest of the world) in terms of either labor, *L*, or natural resources, *NR*, i.e., $K_a/L_a > K_b/L_b$ and $K_a/NR_a > K_b/NR_b$. Because of the equilibrium conditions that the consumption ratios of these factors must be equal and the value of a country's indirect exports of factors must equal its indirect imports of factors, it follows that country *A* must in effect export capital to country *B*. In a factor-price equalization model, this in turn implies that country *A*'s exports will be capital-intensive compared to its imports.

[13] See the book by Hal Lary and the articles by Gordon Philpot and Merle Yahr.

[14] The possible lack of global univalence between factor prices and commodity prices when goods are produced with specific natural resources as well as with capital and labor should be further investigated. This corresponds to factor-intensity reversals in the two factor-two commodity case. James Ford (p. 60) raises this point.

Demand Bias

Demand bias is invariably cited as a possible explanation of the Leontief paradox but no writer has strongly argued that this is the major explanation. Indeed it is now usually accepted (see, for example, the article by Arthur Brown and the analysis by Travis, pp. 105–10) that, if final demand in the United States is factor-biased, the bias is towards labor-intensive rather than capital-intensive goods, because of the operation of Engel's law.

Tariffs and other Distortions

The argument that various tariff and nontariff trade-distorting measures account for Leontief's results has been expounded most cogently by Travis. He arrives at this conclusion after carefully showing the correctness of Leontief's test of the Heckscher-Ohlin theory and then arguing that it is highly unlikely that the failure of the various assumptions of the Heckscher-Ohlin theory to hold (other than the free trade one) could account for Leontief's results.[15]

Another explanation stressing the importance of market imperfections is Diab's suggestion (pp. 53–56) that commodities produced abroad by American corporations or their subsidiaries and with the aid of American capital, know-how, and highly skilled technicians and managers should be regarded as part of *U.S.* internal trade rather than imports. Since a large part of this production consists of capital-intensive, natural resource products (especially minerals), the paradox might well be reversed if these were excluded from the trade pattern. However, as Travis points out (pp. 110–11), for this argument to be valid it is necessary to explain why American capital, once overseas, does not move into labor-intensive industries in foreign countries.

Although Travis seems to believe this point cannot be explained staisfactorily, there is considerable evidence in the literature on economic development supporting the view that the immobility of foreign capital and know-how between the export and domestic sectors of less developed countries is a real phenomenon and is based on economic factors. Consider, for example, why foreign funds have in the past flowed mainly into export-oriented, natural resource industries in the less developed countries or in tertiary lines that serve to support these industries. Part of the explanation is that there usually is better knowledge in the developed countries concerning profit opportunities in the developing countries with regard to natural resource industries compared to most other products. Because of the generally lower supply elasticities in developed countries for natural resource products than for commodities produced mainly with capital and labor, there tends to be a greater upward pressure on the prices of those natural resource products that are significant inputs into industrial processes than on the prices of most other products, as growth takes place in the advanced countries. This relative price movement alerts investors to the obvious profit opportunities that can be exploited if costs can be kept from rising and thereby leads to a search at home and abroad for new supply-sources as well as for better ways of using existing supply sources. On the other hand, even if highly profitable opportunities exist overseas in product lines outside of the natural resource group, investors are less likely to become aware of them because of the absence of this signaling mechanism.

Other important factors affecting for-

[15] Tariffs can weaken the pattern of indirect factor-trade in a Heckscher-Ohlin model but cannot alone produce paradoxical results. Export subsidies (or some domestic distortion) in lines that intensively use a country's relatively scarce factors are needed to produce these results.

eigners' decisions to invest in less developed countries are the nature of factor supplies in these countries, the size of markets, and the degree of input-complexity of production. Natural resource conditions are often sufficiently favorable to make foreign investment profitable in large scale, primary industries that can supply the large markets of developed countries. However, for products that do not rely heavily on the natural resource factor, production costs usually are too high for exports to be internationally competitive. Labor with very little skill is abundant, but without some training this labor is very inefficient even when used in producing the simplest types of manufactures under modern methods.

A lack of trained labor is less of a barrier to the competitive production of manufactured goods for domestic consumption. However, the costs of establishing and supervising productive units abroad tends to be prohibitive unless the optimum plant size is large. But, in industries where the optimum size of productive units is large, domestic demand usually is too small to support efficient production. Still another factor discouraging foreign investment is the more complex system of input requirements (direct and indirect) for manufactures than for primary products. It is more difficult to finance, coordinate, and fully utilize interdependent investment projects in several as compared to a few industries.

After foreign capital moves into export-oriented, natural resource industries in less developed countries, it does not then flow into domestic industries for the same reasons foreign capital does not move directly into these industries. One difference, however, is that foreign firms located within less developed countries have some advantage over outside investors in ascertaining profit opportunities in other fields. However, there is considerable immobility of capital from such foreign-owned and foreign-directed firms into new product lines involving a very different technology from that used for existing production, especially if the optimum plant size is small. Foreign firms engaged, for example, in oil or copper production will vigorously seek out further profit opportunities in their own product lines, including those that establish additional forward and backward production linkages. But, a lack of interest and knowledge concerning the production and marketing of completely different products tends to offset their proximity advantage.

The flow of direct investment funds into developing countries involves not only an increase in the capital stock of these countries, but also an improvement in technology in the sectors affected. This means that, since foreign capital does not move into very many domestically-oriented industries, the state of technology in these industries remains backward. Thus, the explanation of why many less developed countries export capital-intensive products may rest on the immobility of capital between the export and domestic sectors and the technological disparity between these sectors compared to the same sectors in developed economies.

Although the analysis has dealt thus far with developed countries, it also has some applicability to resource-abundant countries like Canada. *U.S.* investment in natural resource industries in Canada tends to create a greater capital and technological disparity between export and domestically oriented industries than would exist without this investment. However, the experience of recent years has shown that as income and domestic markets grow in such developed countries, direct investment by the United States and other advanced countries takes place in product lines that formerly were mainly imported into these countries and are

characterized by significant scale economies. This seems to occur partly to take advantage of being located near the market and partly as a defensive response to import competing investments by domestic investors. To the extent that *U.S.* investment of this sort is in product lines that are more capital-intensive than other *U.S.* exports, the Leontief paradox tends to be reinforced.

One other matter that should be considered before presenting the empirical results of testing some of the different trade hypotheses is whether the Heckscher-Ohlin proposition should hold with respect to each pair of countries.[16] Given a pure Heckscher-Ohlin-Samuelson model where all goods are traded, the number of products exceeds the number of factors, and factor-price equalization is achieved, the answer is that the proposition need not hold on a bilateral basis. When the number of commodities is greater than the number of productive factors, the precise distribution of world production and trade is indeterminate with a particular distribution of productive factors among countries and a given set of factor prices.[17] It is not necessary, for example, for the most capital-abundant country to export a larger proportion of the total exports of the most capital-intensive product than a less capital-abundant country or indeed to export it at all. Within the limits set by factor prices, the actual pattern of intercountry production of any traded commodity depends upon a host of complex factors related to different historical rates of development. What is required in the Heckscher-Ohlin theory is simply that the capital-labor ratio of a capital-abundant country's total exports be greater than the capital-labor ratio of its imports. It is quite possible for this relationship to hold with regard to a country's total trade but not with respect to its trade with a particular country.[18] As Hodd (p. 22) has pointed out, preventing complete factor-price equalization by introducing transport costs into the two-factor model causes the Heckscher-Ohlin proposition to hold bilaterally as well as multilaterally, but this bilateral relationship can again break down when the model is complicated by additional factors, e.g., natural resources, that are complementary to one of the other factors, e.g., capital. Since Vanek and more recently Lawrence Weiser found evidence of a complementary relationship between capital and natural resources at least for the United States, the several empirical studies that have revealed inconsistencies in the factor-content of trade between two countries and the relative factor-endowment pattern of the two countries should not be regarded as providing evidence that necessarily runs counter to the Heckscher-Ohlin theory as it is now generally formulated.

II. Testing the Heckscher-Ohlin Theory and Other Trade Hypotheses

The major results of retesting the Heckscher-Ohlin hypothesis for the United States, using 1962 trade figures and 1958 capital, labor, and intermediate-input data are presented in Tables 1, 2, and 3. Table 1 presents factor-content (direct and indirect) ratios[19] that compare representa-

[16] Bhagwati (pp. 175–76) raises this issue in his survey article and terms the lack of analysis on this point a serious deficiency in trade theory.

[17] The net factor-trade balance is, of course, the same in these circumstances.

[18] See Appendix A for a geometric illustration of such a case. It is also not necessary under the Heckscher-Ohlin theory for the ratio of capital per worker in export industries to capital per worker in import-competing industries to be higher for a capital-abundant country than some less capital-abundant economy.

[19] In testing the relationship between relative factor supplies and the factor content of trade, some writers (Keesing (1965) and Waehrer) compute only the direct factor content of exports and import replacements on the grounds that most intermediate inputs can be imported instead of produced domestically. This procedure confuses an *ex post* test of an equilibrium trade position

tive bundles of import competing with export products;[20,21] Table 2 shows the distribution of the labor force by broad occupational groups; and Table 3 gives different regression estimates in which net adjusted trade balances[22] of the various trading industries in the 1958 input-output table are made a function of different economic factors that allegedly influence the commodity pattern of trade.

One important result of the test is that the Leontief paradox still holds.[23] The ratio of capital per man-year in import-competing versus export production is 1.27[24] compared to the ratios of 1.30 and 1.06 that Leontief obtained for the 1947 and 1951 trade patterns, respectively.[25] Furthermore, in the various stepwise multiple regressions that were performed, the capital-labor ratio always entered first with a statistically significant negative sign as the single variable that best "explained" the trade pattern. However, if natural resource products are excluded, the

to determine if the pattern of trade is consistent with the Heckscher-Ohlin theory with such exercises as predicting or planning for the detailed nature of a country's trade pattern, given its factor endowment and a set of international commodity prices. For the latter purpose the investigator must consider the possibility with respect to any possible export product that national income may be made greater by importing intermediate products rather than producing them domestically. Consequently, the optimum position may well be one where many intermediate inputs involved in trade are not produced locally. Lary's study of the potentialities for exports of manufactures in developing nations in which he analyzes only direct value-added ratios illustrates a problem where the use of direct factor-content ratios is the proper procedure. However, given a particular equilibrium pattern of trade, it is necessary to include both the direct and indirect labor and capital involved in producing exports and imports in order to determine a country's net trade balance in factor services via trade in commodities. If only direct coefficients are used, it is possible to conclude, for example, that a capital abundant country exports labor services and imports capital services when in fact it does the opposite. The direct coefficient test thus would erroneously infer that the Heckscher-Ohlin hypothesis failed to hold.

[20] The representative export and import-competing bundles do not include any services but instead are composed entirely of traded commodities.

[21] Intermediate products imported and then reexported in the form of other products as well as imports containing intermediate inputs that were exported and then reimported should, of course, be excluded in calculating net factor flows. The Leontief method does in fact accomplish this since, for example, foreign produced intermediates that are imported and reexported are counted by the Leontief method on both the import and export side and thus net out in subtracting the factor services involved in exports from those involved in imports. If capital-labor ratios of exports and imports are compared, an incorrect ratio will be obtained but the error factor will not effect whether the quotient is above or below unity—which is the main purpose of the calculation.

[22] Each industry's exports and competitive imports were adjusted by multiplying their respective share of total exports and competitive imports by one million dollars. An industry's net trade balance is the difference between these adjusted values for exports and competitive imports.

[23] Gary Hufbauer also obtains this result in his study of *U.S.* trade in manufactures.

[24] An estimate of this capital-labor ratio was also made in which transportation services, travel (weighted by an average of the capital-labor ratios for hotels and personal services, amusements, and miscellaneous manufactures) and other private services (weighted by an average of the capital-labor ratios for communications and radio and TV broadcasting) were included in the export and import competing bundles. The capital-labor ratio for imports rose to $18,300 and that for exports to $15,000. The import ratio divided by the export ratio was, therefore, 1.22.

[25] As Travis (pp. 98–99) has noted, the exclusion of noncompetitive imports from the *U.S.* import bundle because of the nonavailability in the United States of certain natural resources required for their production could conceivably result in an incorrect inference concerning the Heckscher-Ohlin hypothesis from a Leontief-type test. This would be the case if the production of noncompetitive imports was so highly labor-intensive (and would be so in the United States had the specific natural resources been available) that the capital-labor ratio of total imports, in contrast to just competitive imports, was less than the capital-labor ratio of exports. However, on the basis of a rough survey of the capital-labor ratios for noncompetitive imports produced abroad and given the fact that these imports constitute only about 8 percent of total *U.S.* commodity imports, it appears to be extremely unlikely that the labor intensity of noncompetitive imports could be so high as to account for the Leontief paradox. Actually, the capital-labor ratio calculated in the paper for competitive imports is so high that the capital-labor ratio for the 8 percent of imports (or 6 percent if traded services are included) which are noncompetitive would have to be negative in order to make the overall capital-labor ratio for imports even equal to the capital-labor ratio for exports.

TABLE 1—FACTOR REQUIREMENTS (DIRECT AND INDIRECT) PER MILLION DOLLARS OF *U.S.* EXPORTS AND COMPETITIVE-IMPORT REPLACEMENTS, 1962

	Imports	Exports	Import/ Export Ratio
Net Capital			
All Industries	$2,132,000	$1,876,000	1.14
Excl. Agriculture	1,806,000	1,403,000	1.29
Excl. Natural Resource Products[a]	1,259,000	1,223,000	1.03
Gross Capital			
All Industries	$2,393,000	$2,196,000	1.09
Excl. Agriculture	2,083,000	1,777,000	1.17
Excl. N. R.	1,582,000	1,599,000	.99
Labor (man-years)			
All Industries	119	131	.91
Excl. Agriculture	100	109	.92
Excl. N. R.	106	107	.99
Net Capital-Labor			
All Industries	$18,000	$14,200	1.27
Excl. Agriculture	18,100	12,800	1.41
Excl. N R.	11,900	11,500	1.04
Average Years of Education of Labor			
All Industries	9.9	10.1	.98
Excl. Agriculture	10.2	10.6	.96
Excl. N. R.	10.3	10.7	.97
Average Costs of Education of Labor			
All Industries	$10,300	$10,500	.97
Excl. Agriculture	11,000	11,900	.92
Excl. N. R.	11,200	12,200	.92
Net Capital Plus Total Cost of Education÷Labor			
All Industries	$28,300	$24,700	1.14
Excl. Agriculture	29,100	24,700	1.18
Excl. N. R.	23,100	23,700	.97
Average Earnings of Labor			
All Industries	$4,570	$4,660	.98
Excl. Agriculture	5,050	5,460	.92
Excl. N. R.	5,030	5,400	.93
Proportion of Engineers and Scientists			
All Industries	.0189	.0255	.74
Excl. Agriculture	.0230	.0352	.65
Excl. N. R.	.0228	.0369	.62
Scale Index			
All Industries	51	56	.91
Excl. Agriculture	55	66	.83
Excl. N. R.	57	67	.85
Unionization Index			
All Industries	59	62	.95
Excl. Agriculture	65	72	.90
Excl. N. R.	65	71	.92
Concentration Index			
All Industries	39	40	.98
Excl. Agriculture	42	46	.91
Excl. N. R.	41	46	.89
Proportion of Labor with			
0–8 years of education	.39	.37	1.05
9–12 years of education	.49	.50	.98
13+ years of education	.12	.13	.92

(Footnote [a] will be found on next page.)

capital-labor ratio falls to 1.04 when capital is measured on a net basis and to 1.00 when capital is measured in gross terms.

The hypothesis that export production involves higher skill requirements than import competing production also receives support, as the figures on average earnings, average years of education, and average costs of education indicate. A crude measure of the amount of physical and human capital used in export versus import-competing production was calculated by combining the data on physical capital and the costs of education.[26] As Table 1 indicates, adding this measure of human capital to the physical capital figure is not sufficient to reverse the Leontief results for all industries combined but does reverse it when natural resource industries are excluded.

Classifying the labor force involved in export and import competing production by levels of education and by various occupational groups further brings out the importance of the skill factor in explaining *U.S.* trade. The educational breakdown

[26] To obtain direct education costs, the figures on years of education from the 1960 sample census, supplemented with data on school retention rates, were multiplied by the 1956 cost figures determined by Theodore Schultz (p. 34). Estimates of foregone earnings were added to these direct costs to obtain total education costs. However, no measure of accumulated interest costs is included in the estimate nor does it include any on-the-job training costs.

[a]Natural resource products were arbitrarily defined as all agricultural and mining industries (1–10); tobacco manufactures (15); lumber and wood products (20); petroleum refining (31); and primary nonferrous metals manufacturing (38). The list is roughly similar to the one used by Leontief except that petroleum refining is added and non-livestock agricultural products are included. One could argue quite persuasively that other industries should also be included.

Sources: The coefficients of total requirements (direct and indirect) per dollar of delivery to final demand were taken from U.S. Department of Commerce. The employment figures used to calculate the 1958 direct labor coefficients for the 79 industries covered in the study were furnished by Jack Alterman of the Bureau of Labor Statistics.

The 1958 capital coefficients for industries 12–64 were obtained by reclassifying data given in *Census of Manufactures, 1958*. Net capital is the sum of net book value, work in progress, materials, and finished product inventories. Finished product inventories were adjusted to purchasers prices and, by utilizing the transaction matrix of the input-output table, were distributed to the industries using them. The gross capital coefficients are based on gross book value rather than net book value. The coefficients for the non-manufacturing sectors are based on a wide variety of sources. They include the two basic Leontief articles (1953 and 1956); the study by Leontief et al.; Bert G. Hickman; John W. Kendrick; and Daniel B. Creamer, Sergei P. Dobrovolsky and Israel Borenstein.

Export and import data for the 60 commodity sectors in which trade occurred in 1962 are from *Exports and Imports as Related to Output*. Values of exports and imports for 1962 were adjusted to 1958 prices by deflators that in the case of the manufacturing sectors (13–64) were provided by the Office of Business Economics, Department of Commerce and that for mining and agriculture were obtained from the *1962 Minerals Yearbook* and *Wholesale Prices and Price Indexes*, 1962. Imports were multiplied by ratios of landed value to foreign port value and exports by ratios of producer value to export value in order to make them comparable to the producer-value figures of the input-output table. These ratios were provided by the Office of Business Economics, Department of Commerce.

The Bureau of the Census, *1/1000 Sample of the Population of the United States, 1960* was available in data tape form from the Social Systems Research Institute computation library at the University of Wisconsin. Using the industrial classification system employed by the Census, it is possible to arrange the labor force covered in the census into the same industry groups adopted for the 1958 input-output table. Characteristics relating to years of education, occupations, and earnings were then determined from the data tape for these individuals.

The direct scale and unionization ratios were adapted from Leonard Weiss. The direct scale index for each industry is based on the percentage of employees in establishments with 250 or more employees, and the direct unionization index represents the percentage of an industry's production workers employed in plants where a majority of the workers are covered by collective bargaining contracts. The direct concentration figures, also from the paper by Weiss, are four-firm concentration ratios, adjusted for the local or regional character of certain industries. The row vector of direct ratios for each of the three variables was postmultiplied by the inverse matrix, and weighted averages of the direct and indirect requirements for these characteristics then obtained for each industry by dividing the resulting row vector by the appropriate column sums of the inverse matrix.

TABLE 2—DISTRIBUTION OF LABOR FORCE BY SKILL GROUPS, PER MILLION DOLLARS OF EXPORTS AND COMPETITIVE-IMPORT REPLACEMENTS, 1962
(in percentages)

	Imports	Exports	Import/ Export Ratio
(A) *Six Skill Groups*			
I. Professional, technical and managerial	12.0	12.5	.96
II. Clerical and sales	15.2	15.1	1.01
III. Craftsmen and foremen	14.9	15.4	.97
IV. Operatives	30.4	25.1	1.21
V. Laborers (nonfarm) and service	10.3	7.5	1.37
VI. Farmers and farm laborers	17.2	24.4	.70
	100.0	100.0	
(B) *Eleven Skill Groups*			
I. Professional and technical	5.7	6.7	.85
II. Managerial, except farm	6.3	5.8	1.09
III. Craftsmen and foremen	14.9	15.4	.97
IV. Sales	4.4	4.1	1.07
V. Clerical	10.8	11.0	.98
VI. Operatives	30.4	25.1	1.21
VII. Laborers, except farm	6.9	4.3	1.60
VIII. Service, except private household	3.1	2.9	1.07
IX. Farmers and farm managers	11.2	15.8	.71
X. Private household workers	.3	.3	1.00
XI. Farm laborers and foremen	6.0	8.6	.70
	100.0	100.0	

Sources: See Table 1.

indicates that the proportions of individuals with 9–12 years of education and especially with 13 or more years of education are higher in export than in import competing production, whereas the share of those with only 0–8 years of education is higher on the import side. As the occupational figures (Table 2) show, farmers and farm laborers, who are among the least educated occupational groups, are considerably more important in export than import competing production. However, nonfarm laborers and operatives, who are also at the lower end of the educational attainment scale, are sufficiently more important in import competing production compared to export activities to make the proportion of the labor force as a whole with only a primary school education more significant in import competing than export production. The other occupations that stand out as more significant on the export than import competing side are professional, technical and managerial employees and craftsmen and foremen. Clerical and sales employees do not differ in their relative importance in export versus import competing activities.

The correlation analysis (Table 3, Equations 1a, 2, and 3) shows that there is a significant positive relationship between the percentage of engineers and scientists, craftsmen, and farmers in an industry and the net world export surplus of the industry. The percentage of operatives and nonfarm laborers have the expected negative signs but the coefficients are not significant. As regression equation 7 indicates, there is also a statistically significant positive relationship between the importance in an industry of those with more than a high school education and the industry's world trade balance.

Research and development activities also show up as being much more important in export output than in import-competing production. The ratio of the R&D costs involved in producing a representative bundle of import-competing versus export commodities, as calculated from the R&D sector in the input-output table, is .66.[27] The ratio of the number of engineers and natural scientists engaged in import-competing versus export activities

[27] The R&D sector in the input-output table includes, however, only research and development performed for sale and thus excludes R&D performed within a company.

TABLE 3—REGRESSION EQUATIONS RELATING SELECTED ECONOMIC CHARACTERISTICS BY INDUSTRY TO *U.S.* WORLD AND BILATERAL TRADE BALANCES BY INDUSTRY

	Independent Variables												
Dependent Variable (in dollars)	Capital-Labor	Cap. Plus Costs of Educ.-Labor	Percentage of the Labor Force in Various Skill Groups: Eng. & Sc.	Rest of I	II	III	IV	V	VI	Scale Index	Union-ization Index	Conc. Index	R^2
1a. Adjusted Exports Minus Adj. Imports[a] (World)	−1.37	—	7011	−1473	71	1578	−248	−761	845	−421	343	—	.44
	*−4.35***		*2.13**	*− .69*	*.06*	*1.96**	*− .79*	*− .80*	*3.81***	*−1.25*	*1.11*		*3.85***
1b. *X-M* (Canada)	−1.38	—	−4608	1778	512	1127	−335	−2771	546	302	2	—	.24
	*−2.39**		*− .76*	*.45*	*.22*	*.76*	*− .58*	*−1.81**	*1.35*	*.49*	*.00*		*1.52*
1c. *X-M* (W. Europe)	− .27	—	8185	1397	−1040	907	34	−958	1095	−983	389	—	.32
	−.68		*1.98**	*.52*	*− .66*	*.90*	*.09*	*− .91*	*3.93***	*−2.32**	*1.00*	—	*2.31**
1d. *X-M* (Japan)	− .06	—	8748	2242	−930	1010	−718	−2429	1333	−1381	1047	—	.42
	− .13		*1.74**	*.68*	*− .48*	*.82*	*−1.49*	*−1.90**	*3.93***	*−2.08***	*2.22**		*3.51***
1e. *X-M* (LDCs)	−2.73	—	−768	−1373	−948	2701	−315	435	659	255	220	—	.51
	*−5.80***		*− .16*	*− .43*	*− .51*	*2.25**	*− .67*	*.35*	*1.99**	*.51*	*.48*		*4.99***
1f. *X-M* (Others)	−1.11	—	27496	−13339	3761	2582	− 15	3041	− 402	− 201	− 506	—	.37
	−1.49		*3.54***	*−2.64**	*1.27*	*1.36*	*− .02*	*1.54*	*− .77*	*− .20*	*− .70*		*2.84***
2. *X-M*[a] (World)	−1.20	—	5789	−1145	− 84	1603	−310	−728	854	—	295	−478	.44
	*−3.98***		*1.97**	*− .52*	*− .07*	*1.97**	*−1.01*	*− .86*	*3.75***		*.97*	*−1.03*	*3.76***
3. *X-M*[a] (World)	—	−1.36	7631	−1156	298	1663	−184	−620	928	−406	345	—	.45
		*−4.38***	*2.30**	*− .54*	*.24*	*2.05**	*− .58*	*− .74*	*4.05***	*−1.21*	*1.12*		*3.88***

Dependent Variable	Constant	Capital-Labor	Cap. Plus Costs of Educ.-Labor	Percentage of Labor Force With Various Yrs. of Educ.: 1–8 yrs.	9–12 yrs.	13+yrs.	Percent Sc. & Eng.	Av. Costs of Educ.	Av. Yrs. of Educ.	Scale Index	Union-ization Index	Conc. Index	R^2
4. *X-M* (World)	65256	−1.09	—	—	—	—	10080	−6.64	—	−319	259	—	.25
	1.66	*−3.29***					*2.25**	*−1.50*		*− .86*	*.90*		*3.45*
5. *X-M* (World)	108312	−1.12	—	—	—	—	—	—	−10728	−305	249	—	.23
	1.16	*−3.32***							*−1.06*	*− .81*	*.85*		*3.16***
6. *X-M* (World)	17246	—	−1.04	—	—	—	4274	—	—	—	81	−157	.21
	1.02		*−3.26***				*1.67*				*.31*	*− .31*	*3.60***
7. *X-M* (World)	—	−1.35	—	640	−844	3090	—	—	—	—	—	—	.22
		*−3.86***		*−1.57*	*−1.46*	*2.00**							*3.83***

Sources; See Table 1

[a] The constant term has been suppressed in equations 1, 2, 3, and 7 because of the indeterminancy resulting from the fact that the skill-level and education-level percentages add to unity in each industry.

* and ** indicate 90 percent and 99 percent significant levels for the *t* values (shown in italics) of the regression coefficients and for the *F*-ratio of the squared multiple correlation coefficient.

is .74.[28] Moreover, as already noted, in the regression model used, the percentage or the absolute number of engineers and scientists in an industry appears as a significant variable that is positively correlated with the industry's export surplus. This relationship is especially strong when natural resource products are excluded from the trade pattern.

Another exercise confirming the importance of this variable is the correlation between the percentage change of exports in each industry from 1947 to 1962 and various characteristics of the labor force in each industry, such as their earnings, years of education, a simple skill index, the absolute importance of engineers and scientists as well as general industry characteristics such as the degree of concentration, unionization, and large scale employment.[29] The engineers-scientists variable and the concentration ratio are the only two variables that are significantly related (positively) to the growth of exports. When the same variables are used to explain the percentage change in imports, none comes out to be statistically significant.

Beside determining the trade requirements for engineers and scientists, an estimate was made of the import versus export requirements of top management, i.e., managers, officials, and proprietors earning more than $10,000. Their numbers are larger in export production than in import-competing production but the proportions of the total labor force engaged in top management activities are about the same on the import and export sides.[30]

Two other industry characteristics of special interest are the relative importance of scale economies and the degree of unionization in import-competing versus export production. Steffan Linder has stressed the point that profitable production for home markets is a necessary condition for manufactured products to be potential export products. Consequently, in industries where scale economies are important, the size of the American market may give the United States an export advantage, quite aside from any other factors. The unionization calculation is aimed at the hypothesis that unions may raise wages above their competitive levels and thus act to offset underlying "real" factors that contribute to a country's comparative advantage.[31] A variable reflecting the degree of industrial concentration is also introduced into the analysis, but it is highly correlated ($r=.87$) with the scale index. As Table 1 indicates, the scale factor, the degree of unionization, and the degree of concentration are all more important for export production than for import-competing activities. However, none of these variables turn out to be sta-

[28] This group includes both individuals engaged in research and development as well as those engaged in current production activities. Using data for eighteen industries and direct requirements only, Kenen (1968) compared the relative importance of the two groups in "explaining" trade patterns and obtained ambiguous results. As Keesing (1968 pp. 175–89) had previously shown, for exports alone, the ratio of scientists and engineers engaged in research and development to the total labor force in the industry is statistically significant whereas the proportion of scientists and engineers in non-R&D activities is not. On the other hand, when an industry's net trade balance is taken as the dependent variable, the opposite result is obtained. Consequently, in view of these results and also because of the similarity between the results obtained in this study from using R&D expenditures and the number of scientists and engineers, it seems best to regard the engineers-scientists variable used here as both a skill measure and a proxy for R&D activities that result in new and improved products.

[29] In this exercise the 1947 trade data were classified on the basis of the industrial breakdown in the 1958 input-output table. The various economic characteristics of the industries pertain to the period around 1960.

[30] Two other results that may be of interest are that there is no difference between import-competing and export production in the average age of the workers or the proportion who are white.

[31] Even assuming that unions do raise wages above competitive levels, this hypothesis depends on the assumption that the resulting competitive disadvantage is not offset by the same force operating in other countries.

tistically significant in the regression analyses of total trade.

In order to indicate the effect of import duties as well as some of the main nontariff trade barriers on the capital-labor ratio employed in import-competing production, import demand elasticities were assigned to the various trading industries in the input-output table and a new per million dollar bundle of competitive-imports was then determined under the assumption that the average duty (or the ad valorem duty-equivalent of the nontariff barrier) in each industry was reduced to zero.[32] The fact that the capital-labor ratio with the new commodity-composition of imports is about 5 percent lower than the ratio computed with the actual import bundle confirms Travis's contention that tariffs operate in the direction of the Leontief paradox.[33] Furthermore, the commercial policies of other countries probably tend to reduce the average capital-labor ratio in export production below its free trade level. A removal of all trade-distorting measures might confirm the expectation of the Heckscher-Ohlin model for the United States, but, based on my own study of tariff and nontariff barriers to trade, I do not think that this would be the case. However, further empirical study of this subject is very much needed.

The effect of an increase in income levels on the capital-labor ratio of import competing production was also estimated by assigning appropriate income elasticities of import demand to various commodity groups.[34] The capital-labor ratio required to produce the increment in imports associated with an increase in income is \$17,750 or slightly less than the \$18,000 average for the 1962 bundle of competitive imports. Thus, if the *U.S.* commodity structure of income elasticities of import demand is typical of the pattern for the rest of the world towards the United States, then demand differences related to income differences among countries are not a factor that tends to account for the Leontief paradox.

In addition to determining the factor content of a representative bundle of *U.S.* exports to all countries as a whole and a representative bundle of competitive imports from the rest of the world, the Heckscher-Ohlin hypothesis was tested with respect to *U.S.* trade vis-à-vis Western Europe, Japan, Canada, and less developed countries, and an all other group (mainly Oceania).[35] As previously noted, the assumptions necessary for the Heckscher-Ohlin proposition to be logically true with regard to a country's total

[32] Using the study by R. J. Ball and K. Marwah, industries listed in the input-output table were divided into five groups and assigned the following import-demand elasticities: crude foodstuffs −.46; manufactured food-stuffs −2.39; crude materials −.38; semi-manufactured products −1.64 and manufactured goods −4.04. Tariff rates for 1962 were obtained by dividing the calculated import duty for an industry by the value of its imports. The nontariff barriers included were the quotas on agricultural products, cotton textiles, and petroleum as well as the American Selling Price system of valuing certain chemicals. See Baldwin (p. 163). In estimating the price effect of the cuts in the degree of protection, it was assumed that the elasticity of foreign import-supply was infinite for all commodities.

[33] The indirect effect that reducing duties on products used as intermediate inputs has in increasing domestic production and thus reducing imports was not taken into account in estimating the new import bundle.

[34] The income elasticities as well as the commodity classification employed in making the estimate are from the Ball-Marwah article. The particular income elasticities used were: crude foodstuffs .49; manufactured foodstuffs .96; crude materials .87; semi-manufactures 1.22; and manufactured goods 2.47.

[35] Western Europe consists of all European members of OECD; the less developed countries are composed of other Asia (Asia except for Japan and China Mainland), Africa, and other America (Americas excluding the United States and Canada); and the all other group is made up of Eastern Europe, other Europe, China Mainland, and Oceania. The data on which these regional trade patterns are based are much less detailed than those from which the world trade pattern is derived. Furthermore, the fact that no products are excluded from the export side on the basis of being noncompetitive is a more serious drawback for bilateral analyses than for the analysis of *U.S.* trade with the rest of the world.

TABLE 4—FACTOR-CONTENT RATIOS FOR *U.S.* BILATERAL TRADE PATTERNS WITH SELECTED REGIONS, 1962

	Import/Export Ratios[a]									
	Canada		Western Europe		Japan		LDCs		Others	
Economic Characteristic	All Sectors	Excl. N.R.	All Sectors	Excl. N.R.	All Sectors	Excl. N.R.	All Sectors	Excl. N.R.	All Sectors	Excl. N.R.
Net Capital	1.25	1.02	.83	1.01	.77	1.03	1.44	1.05	1.52	1.09
Gross Capital	1.16	.99	.87	.99	.83	1.03	1.28	.93	1.40	1.04
Labor	.89	.88	.95	1.07	1.05	1.23	.81	.92	1.07	.82
Net Capital-Labor	1.41	1.15	.87	.93	.73	.84	1.78	1.14	1.42	1.33
Average Years of Education	.99	1.01	1.02	.98	1.01	.95	.98	.96	.92	.95
Average Costs of Education	.97	.99	1.05	.94	1.03	.88	.96	.89	.81	.87
Net Capital Plus Cost of Education+Labor	1.19	1.07	1.19	.89	1.22	.87	1.24	1.01	1.64	1.06
Average Earnings	.97	.91	1.05	.94	1.04	.88	.92	.81	.80	.90
Proportion of Engineers and Scientists	.82	.90	1.04	.74	.94	.64	.75	.47	.29	.37
Scale Index	.77	.86	1.27	1.07	1.28	1.07	.70	.68	.58	.62
Unionization Index	.86	.95	1.20	1.05	1.14	.99	.77	.81	.77	.83
Concentration Index	.83	.85	1.20	1.05	1.10	.96	.82	.78	.75	.77
Skill Group I[b]	.96	.94	1.10	.87	1.08	.77	.93	.83	.66	.81
II	.89	.96	1.18	.96	1.16	.88	.95	1.04	.75	1.13
III	.91	.89	1.20	.92	1.12	.81	.83	.71	.62	.89
IV	1.04	1.07	1.42	1.14	1.67	1.32	1.03	1.23	.82	1.12
V	1.60	1.13	1.21	1.02	1.44	.95	1.12	1.06	1.24	1.44
VI	.82	—	.46	—	.24	—	1.11	—	.24	—

Sources: Bilateral trade data are from OECD, *Foreign Trade Statistical Bulletins.* See Table 1 for sources of other data.

[a] Import/Export ratios are computed in the same manner as in Table 1

[b] The skill groups are the same as in Table 2.

trade do not imply that the theory must hold on a bilateral basis. However, a regional analysis is useful in revealing additional information on the factors influencing the commodity pattern of *U.S.* trade. The results of these tests (Table 3, equations 1b–1f and Table 4) are that the Leontief result does not hold with respect to either *U.S.*-Western European or *U.S.*-Japanese trade but does exist with respect to trade between the United States and Canada, the United States and less developed countries, and the United States and all other countries. The latter three groups of countries represent regions that are relatively abundant in natural resources. In view of the strong complementary relationship between capital and certain natural resources and the previously made point concerning the international flow of *U.S.* capital, technology, and top management into export-oriented, natural resource industries in foreign countries, the results with regard to Canada, the less developed countries, and Oceania are not unexpected.[36] The trade

[36] As Table 4 indicates, imports from these regions are more capital-intensive than exports to them even when the list of natural resource industries are excluded from the calculations and when human capital is added to physical capital. However, the nature of much of the remaining trade is still greatly influenced by transportation and technical processing considerations that favor location of production in these areas. Such is the case, for example, in the very important food sector (14) where imports from the *LDCs* are dominated by cane sugar and imports from the all other group by meat as well as for the paper industry (24) where imports of pulp and newsprint from Canada are large. When these two industries are also excluded from the factor content calculations, the ratio of physical plus human capital to labor is lower in import competing than export production for all regions, and the ratio of physical capital

patterns with Western Europe and Japan are not as heavily influenced by imports of natural resource products nor is direct foreign investment as important in these regions relative to domestic capital accumulation. Thus, relative domestic supplies of capital and labor play a more important role in determining the trade structure between the United States and these regions.

Although the United States exports comparatively capital-intensive products to Western Europe and Japan, the capital-labor variable does not show up as statistically significant in the multiple regression analysis with respect to these areas. Other factors appear to be more important as determinants of the trade patterns between the United States and these regions. In particular, there is a significant positive relationship between the percentage of engineers and scientists employed in an industry and the industry's net export balance with respect to each of these two regions (as well as for the all other group). The scale variable also shows up as significant for these two regions. Rather surprisingly, the sign of the scale coefficient is negative (largely because of the large export surplus of agricultural products to Western Europe and Japan).[37]

General measures of skill and human capital such as average costs of education, average years of education, and average earnings were found to be statistically significant only in the case of *U.S.*-Japanese trade.[38] However, the percentage of employees with 13 or more years of education required to produce a given value of output in each industry is significantly correlated (positively) with an industry's net trade balance between the United States and Western Europe, the United States and Japan, and the United States and the all other group. A similar positive correlation holds for *U.S.*-Western European trade with regard to the proportion of the labor force educated 8 years or less, whereas a significant negative relationship exists for *U.S.*-Western European and *U.S.*-Japanese trade with regard to those receiving 9–12 years of education. Dividing the labor force into broad occupational groups further reveals the importance of various types of labor skills in a manner generally consistent with what one would expect from a factor-proportions approach. The significance of the engineers-scientists variable has already been mentioned. The number[39] and percentage of unskilled (nonfarm) workers employed in an industry are significantly correlated (negatively) with the industry's export surplus in the cases of Japan and Canada, and the number (though not percentages) of semi-skilled workers enters significantly with the same negative sign for Japan. Furthermore, the number and percentage of skilled workers in an industry is significantly correlated in a positive manner with the industry's net balance of trade between the United States and the less developed countries.

III. Conclusions

The preceding analysis strongly supports the view that a straightforward application of a two-factor (capital and labor) factor-proportions model along Heckscher-Ohlin lines is inadequate for understanding the pattern of *U.S.* trade. Not only is the sign of the capital-labor ratio opposite from what would be ex-

to labor is lower in import competing than export production for all regions except the less developed countries.

[37] When natural resource products are eliminated, the scale factor is not statistically significant for Western Europe or Japan.

[38] These results as well as those that are discussed in the next two sentences are not reported in Table 3.

[39] A regression equation (not shown in Table 3) was estimated in which the independent variables were the capital-output ratio for each industry and the number of employees in each of the six skill groups.

pected from the model but it is statistically significant in this unexpected direction. What this negative sign seems to reflect is, as Vanek and others have suggested, that there is a strong complementarity between certain natural resources—many forms of which are relatively scarce in the United States—and physical capital.[40] The regional breakdown indicates, for example, that the source of the paradox is the pattern of *U.S.* trade with Canada, the less developed countries, and the all other group—all of which export a significant volume of natural resource products to the United States. When various natural resource products are eliminated from the factor-content calculations, the overall ratio of capital per worker in import-competing production to capital per worker in export production drops from 1.27 to 1.04. Omitting natural resource industries from the regression data also eliminates the capital-labor ratio as a statistically significant variable. Moreover, in the remaining group there are still some important industries in which the costs of transporting natural resource products required as inputs are relatively high and whose location, therefore, tends to be near the source of the natural resources that are indirectly required for production.

As previously noted, the complementary relationship between physical capital and natural resources need not be offset by *U.S.* exports of goods that are even more capital-intensive than natural resource products or by other imports of a highly labor-intensive nature, provided the various assumptions of a simple Heckscher-Ohlin model with respect to capital mobility, homogeneity of the labor supply, commercial policy, and technological parity do not hold. Evidence indicating that the capital mobility assumptions do not hold is available in the economic development literature whereas data consistent with the position that the labor-supply, technology, and commercial-policy assumptions of the traditional model do not hold have been presented in this paper. It seems clear from the preceding analysis that the relatively abundant supply of engineers and scientists is an important source of the United States' comparative-advantage position, especially as far as trade in manufactures is concerned. This abundance of highly trained labor gives the United States an export advantage in products requiring relatively large amounts of such labor.[41] Probably of even more importance is the fact that a significant part of this labor group is engaged in research and development activities. Even those working directly in production facilitate the development of product improvements. Thus, in product lines where the technological opportunities for product improvements are favorable, the use of engineers and scientists for research and development activities fosters temporary *U.S.* trade advantages based on technological differences rather than on relative factor proportions. Just how to weigh the relative importance of these two aspects of the engineers-scientists variable cannot be determined from this study but what evidence there is suggests that both are significant in influencing the pattern of *U.S.* trade.[42]

The relative supplies of certain other types of labor skills also appear to be important determinants of the structure of *U.S.* commodity trade. As would be expected in a Heckscher-Ohlin model with

[40] As Seiji Naya (p. 567) has pointed out, this complementary relationship has a more general applicability among countries when agricultural products are excluded from the list of natural resource products.

[41] In most industries engineers and scientists make up only a small fraction of the labor force. However, in 1960 the proportion was between 5 and 10 percent for thirteen of the sixty input-output industries in which international trade took place.

[42] Kenen (1968) concludes from his analysis of this matter that an eclectic approach is still in order.

several types of labor, the United States not only indirectly exports professional and technical labor but also skilled craftsmen and foremen. Furthermore, we indirectly import semi-skilled and unskilled (nonfarm) labor, both of which are usually considered to be comparatively scarce in the United States.

General measures of human capital such as earnings, years of education, and costs of education, fail to capture much of the explanatory power that is given by a breakdown into levels of educational attainment or into traditional skill groups. Part of this may be due to the omission of on-the-job training from the general education variable used in the study. The positive correlation between net exports and average years of education is also weakened by the large export surplus in the agricultural sector, where formal educational requirements are not only low but where the relatively abundant supply of land in the United States plays a significant export-creating role. The existence of social and economic institutional arrangements that impede equilibrating educational adjustments within the labor force and rapid technological progress that frequently changes educational requirements further tend to diminish the statistical significance of gross measures of the stock of human capital as explanatory variables of the *U.S.* trade pattern. Simple (and imperfect) measures of the degree of scale economies, unionization, and industrial concentration also do not account for the *U.S.* trade pattern in a statistically significant manner. Protection levels in each industry, on the other hand, do seem to influence the pattern and factor-content of trade to an appreciable extent.

The clearest conclusion to be drawn from the study is, as other writers (see articles by Kenen (1968) and Hufbauer) have emphasized recently, that it is necessary to discard simple, single-factor (e.g. capital per worker) trade theories in favor of multi-factor trade models. In particular, the labor force must be divided into various skill groups and the notion of relative differences in human capital taken into account. Other variables, such as natural resource conditions, technological differences, transportation costs, and commercial policies, must be explicitly included in these models. Furthermore, trade theory should take greater account of the degree of difficulty with which productive factors move among sectors within an economy and especially barriers to the flow of factors abroad into various sectors of an economy. Under this more general approach the relative abundance among countries of the factors of production will still occupy an important place in trade theory but a more complex notion of productive factors will be utilized and other considerations will also play important explanatory roles. Moreover, as we improve our predictive powers with these broader trade models, we must devote greater efforts to understanding the processes that determine the nature of the underlying variables affecting trade patterns. This, in turn, should enable us to construct a more fundamental, dynamic theory of international trade.

Appendix A

The point that the Heckscher-Ohlin theorem does not necessarily imply that the Heckscher-Ohlin relationship must hold bilaterally, given a standard model with two or more factors, a greater number of products than factors, and at least three countries, can be illustrated with Figure 1. The points *X*, *Y*, and *Z* represent the labor and capital endowment of countries *X*, *Y*, and *Z*, respectively. For simplifying purposes it is assumed that the equilibrium production pattern is such that Country *X* produces only commodity *A* (*OX* of *A*), Country *Z* produces only commodity *C* (*OZ* of *C*), whereas Country *Y* produces *Ok* of *C* plus *km* of *A*,

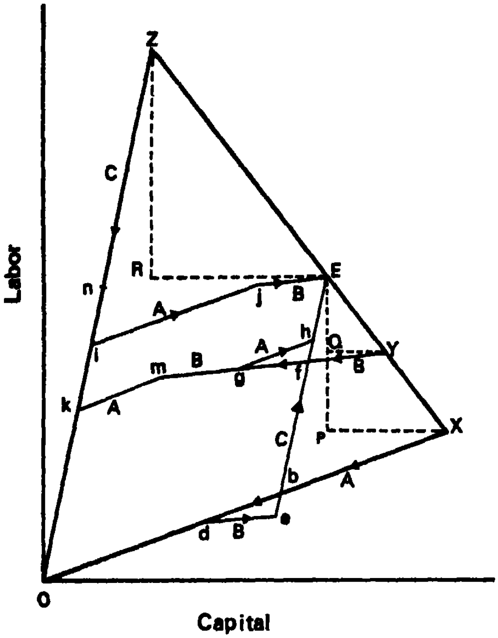


Figure 1

and mY of B. The slopes of the lines representing the production of each product are equal to the equilibrium labor-capital ratios used in producing these goods and are the same in all three countries for each commodity. A given distance along a particular commodity's factor-ratio line also represents the same quantity of the product in all three countries. Given the assumption of identical tastes in all three countries,[43] equilibrium requires that each country consume the three commodities (and two factors) in the same proportions. Again, in order to simplify the figure, the example is constructed so that in the equilibrium situation the absolute quantities of products and factors consumed are actually the same for the three countries after trade takes place.

Country X exports Xd of commodity A ($Xb=ij$ to Country Z and $bd=gh$ to Country Y); Country Y exports Yg of commodity B ($Yf=de$ to Country X and $fg=jE$ to Country Z); and Country Z exports Zi of commodity C ($Zn=eE$ to Country X and $ni=hE$ to Country Y). Thus, for example, Country X exports Xd of A and imports de of B and eE of C. The arrows along these lines indicate the country's trading pattern and show that the country ends up at the consumption point E with od of A, de of B, and eE of C. In factor terms the country trades XP of capital for PE of labor. Country Y, through its commodity trade, in effect exports YQ of capital and imports QE of labor, whereas Country Z exports ZR of labor (equals PE plus QE) and imports RE of capital (equals XP plus YQ). (The slope of line ZEX represents the equilibrium factor-price ratio.) Countries Y and Z also end up at the consumption point E with the same commodity consumption pattern as X.

Although the Heckscher-Ohlin proposition holds in a multilateral sense it does not hold bilaterally. Specifically, although X is more capital-abundant than Y, X's imports from Y are more capital-intensive than its exports to Y.[44]

REFERENCES

K. J. Arrow, H. B. Chenery, B. S. Minhas, and R. M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Statist.*, Aug. 1961, *43*, 225–50.

R. E. Baldwin, *Nontariff Distortions of International Trade*, Washington 1970.

D. S. Ball, "Factor-Intensity Reversals in International Comparison of Factor Costs and Factor Use," *J. Polit. Econ.*, Feb. 1966, *74*, 77–80.

R. J. Ball and K. Marwah, "The U.S. Demand for Imports, 1948–58," *Rev. Econ. Statist.*, Nov. 1962, *44*, 395–401.

P. Bardhan, "International Differences in Production Functions, Trade and Factor Prices," *Econ. J.*, Mar. 1965, *75*, 81–7.

G. S. Becker, *Human Capital*, New York 1964.

J. Bhagwati, "The Pure Theory of International Trade: A Survey," in *Surveys of*

[43] Identical tastes are not necessary for the relationship being illustrated to hold.

[44] As can be seen by comparing the bilateral trade patterns with the slope of the equilibrium factor-price ratio, X has an export surplus with Y that is balanced by an import surplus with Z.

Economic Theory: Growth and Development, New York 1965, *2*, 173–5.

R. Bharadwaj, "Factor Proportions and the Structure of Indo-U.S. Trade," *Indian Econ. J.*, Oct. 1962, *10*, 105–16.

——— and J. Bhagwati, "Human Capital and the Pattern of Foreign Trade: The Indian Case," *Indian Econ. Rev.*, Oct. 1967, *2*, 117–42.

G. Bickel, *Factor Proportions and Relative Price Under CES Production Functions: An Empirical Study of Japanese-U.S. Comparative Advantage*, Stanford 1966.

A. J. Brown, "Professor Leontief and the Pattern of World Trade," *Yorkshire Bull. Econ. Soc. Res.*, Nov. 1957, *9*, 63–75.

D. B. Creamer, S. P. Dobrovolsky, and I. Borenstein, *Capital in Manufacturing and Mining*, Princeton 1960.

P. David and Th. van de Klundert, "Biased Efficiency Growth and Capital-Labor Substitution in the U.S., 1899–1960," *Amer. Econ. Rev.*, June 1965, *55*, 357–94.

M. A. Diab, *The United States Capital Position and the Structure of Its Foreign Trade*, Amsterdam 1956.

P. T. Ellsworth, "The Structure of American Foreign Trade: A New View Examined," *Rev. Econ. Statist.*, Aug. 1954, *36*, 274–85.

J. L. Ford, *The Ohlin-Heckscher Theory of the Basis and Effects of Commodity Trade*, New York 1965.

W. H. Gruber and R. Vernon, "The R & D Factor in a World Trade Matrix," Univ.-Nat. Bur. Comm. Econ. Res., Conference on Technology and Competition in World Trade, New York Oct. 11–12, 1968.

G. Hanoch, "An Economic Analysis of Earning and Schooling," *J. Hum. Resources*, summer 1967, *2*, 310–29.

K. E. Hansson, "A General Theory of the System of Multilateral Trade," *Amer. Econ. Rev.*, Mar. 1952, *42*, 59–68.

B. G. Hickman, *Investment Demand and U.S. Economic Growth*, Washington 1965.

M. Hodd, "An Empirical Investigation of the Heckscher-Ohlin Theory," *Economica*, Feb. 1967, *34*, 20–29.

G. C. Hufbauer, "The Commodity Composition of Trade in Manufactured Goods," Univ.-Nat. Bur. Comm. Econ. Res., Conference on Technology and Competition in International Trade, New York, Oct. 11–12, 1968.

D. Keesing, "The Impact of Research and Development on United States Trade," in P. Kenen and R. Lawrence, eds., *The Open Economy: Essays on International Trade and Finance*, New York 1968, 175–89.

———, "Labor Skills and Comparative Advantage," *Amer. Econ. Rev. Proc.*, May 1966, *56*, 249–55.

———, "Labor Skills and International Trade: Evaluating Many Trade Flows with a Single Measuring Device," *Rev. Econ. Statist.*, Aug. 1965, *47*, 287–94.

———, "Labor Skills and the Structure of Trade in Manufactures," in P. Kenen and R. Lawrence, eds., *The Open Economy: Essays on International Trade and Finance*, New York 1968, 3–18.

J. W. Kendrick, *Productivity Trends in the United States*, Princeton 1961.

P. Kenen, "Nature, Capital and Trade," *J. Polit. Econ.*, Oct. 1965, *73*, 437–60.

———, "Skills, Human Capital and Comparative Advantage," Univ.-Nat. Bur. Comm. Econ. Res., Conference on Human Resources, Madison, Wisc., Nov. 16, 1968.

I. Kravis, "Wages and Foreign Trade," *Rev. Econ. Statist.*, Feb. 1956, *38*, 14–30.

M. Kreinin, "Comparative Labor Effectiveness and the Leontief Scarce Factor Paradox," *Amer. Econ. Rev.*, Mar. 1965, *55*, 131–40.

H. B. Lary, *Imports of Labor-Intensive Manufactures from Less Developed Countries*, New York 1968.

W. Leontief, "Domestic Production and Foreign Trade: The American Capital Position Re-examined," *Proc. of the Amer. Philosophical Soc.*, Sept. 1953, *97*, 332–49.

———, "Factor Proportions and the Structure of American Trade: Further Theoretical and Empirical Analysis," *Rev. Econ. Statist.*, Nov. 1956, *38*, 386–407.

———, "International Factor Costs and Factor Use," *Amer. Econ. Rev.*, June 1964, *54*, 335–45.

——— et al, *Studies in the Structure of the American Economy*, New York 1963.

S. Linder, *An Essay on Trade and Transformation*, New York 1961.

B. S. Minhas, *An International Comparison of Factor Costs and Factor Use*, Amsterdam 1963.

S. Naya, "Natural Resources, Factor Mix, and Factor Reversal in International Trade," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 561–70.

B. Ohlin, *Interregional and International Trade*, Cambridge 1933.

G. Philpot, "Labor Quality, Returns to Scale and the Elasticity of Factor Substitution," *Rev. Econ. Statist.*, May 1970, *52*, 194–99.

K. W. Roskamp, "Factor Proportions and Foreign Trade: The Case of West Germany," *Weltwertschaftliches Archiv*, 1963, *91*, 319–26.

K. Roskamp and G. McMeekin, "Factor Proportions, Human Capital and Foreign Trade: The Case of West Germany Reconsidered," *Quart. J. Econ.*, Feb. 1968, *82*, 152–160.

T. Schultz, *The Economic Value of Education*, New York 1963.

M. Tatemoto and S. Ichimura, "Factor Proportions and Foreign Trade: The Case of Japan," *Rev. Econ. Statist.*, Nov. 1959, *41*, 442–6.

W. P. Travis, *The Theory of Trade and Protection*, Cambridge 1964.

J. Vanek, "The Factor Proportions Theory: The N-Factor Case," *Kyklos*, Oct. 1968, *21*, 749–56.

———, *The Natural Resource Content of United States Foreign Trade, 1870-1955*, Cambridge 1963.

R. Vernon, "International Investment and International Trade in the Product Cycle," *Quart. J. Econ.*, May 1966, *80*, 190–207.

H. Waehrer, "Wage Rates, Labor Skills, and United States Foreign Trade," in P. Kenen and R. Lawrence, eds., *The Open Economy: Essays on International Trade and Finance*, New York 1968.

D. F. Wahl, "Capital and Labour Requirements for Canada's Foreign Trade," *Can. J. Econ.*, Aug. 1961, *27*, 349–58.

L. A. Weiser, "Changing Factor Requirements of United States Foreign Trade," *Rev. Econ. Statist.*, Aug. 1968, *50*, 356–60.

L. Weiss, "Concentration and Labor Earnings," Paper 6405, Social Systems Research Institute, University of Wisconsin, 1964.

M. Yahr, "Human Capital and Factor Substitution in the CES Production Function," in P. Kenen and R. Lawrence, eds., *The Open Economy: Essays on International Trade and Finance*, New York 1968, 70–99.

Organization for Economic Cooperation and Development (OECD), Foreign Trade Statistical Bulletins, Series B, Commodity Trade, Jan.-Dec. 1962, Paris 1963.

U.S. Bureau of the Census, *Census of Manufactures, 1958, 1*, Washington 1961.

———, *Exports and Imports as Related to Output*, Series ES2, No. 5, Washington 1964.

U.S. Bureau of Labor Statistics, *Wholesale Prices and Price Indexes, 1962*, Bull. 1411, Washington 1962.

———, *1/1000 Sample of the Population of the United States, 1960*, Social Systems Research Institute, University of Wisconsin, Madison.

U.S. Department of Commerce, "The Transactions Table of the 1958 Input-Output Study and Revised Direct and Total Requirements Data," *Surv. Curr. Bus.*, Sept. 1965, *44*, 45–49.

U.S. Department of Interior, *1962 Minerals Yearbook, 2*, Washington 1963.

Optimal Restrictions on Foreign Trade and Investment

By FRANZ GEHRELS*

The theory of optimal restriction of international investment, and its interdependence with restriction on trade, has been developed in recent years by G. D. A. MacDougall, A. E. Jasay, M. C. Kemp, T. Negishi, and R. W. Jones (1967). The question introduced by MacDougall was whether, in analogy with trade interference, a borrowing country could alter the terms of lending to its advantage. Jasay considered the same question from the viewpoint of the capital-exporting country. Kemp and Jones went further to consider the interdependence between goods prices and factor prices in a two-country, two-good, two-factor situation; they found expressions for optimal duties and investment taxes, for both debtor and creditor countries. Of interest also is a recent note by V. K. Ramaswami on the related issue of taxing in a descriminatory way the earnings of immigrant labor. This shows that analogous rules can be developed for interference with the international movement of any factor of production—here we shall, however, restrict the discussion to the flow of goods and capital, keeping labor immobile between countries.

It is the purpose of this paper to modify the Kemp-Jones optimizing procedure, bringing it more nearly into line with conventional tariff optimization. This permits easy extension to the n-goods case for both tariff and tax under full optimization (where both duty and tax can be adjusted freely). It also leads to more general answers for the cases of partial optimization dealt with by Jones (where either tax or duty can be varied, but not both). Jones took only the free trade case for the optimal investment tax, and only the free mobility of capital case for the optimal duty. A last problem we deal with is optimal restrictions on trade and borrowing for a country which has an imperfection in the domestic market for labor.

The method used here for finding optimal duties on trade, and taxes on income from foreign investment, is a natural extension of J. de V. Graaff's treatment of duties on several goods. That is to say, one maximizes a social utility function subject to a transformation function and a balance-of-trade constraint, by adjusting quantities of goods traded or the amount of capital invested. Marginal social utility of any good is assumed to match marginal private utility. Because of perfect competition and no externalities, consumer rates of substitution are equal to rates of transformation in production. At the maximum a Pareto optimum exists, because under competitive conditions and independence of tastes, each individual has maximized utility subject to his income constraint. There is no change which would increase any person's satisfaction without reducing that of others.

We make the conventional assumption that there is no foreign retaliation. The

* Professor of economics, Indiana University. This paper was completed during tenure of a research fellowship at the Institute for International Economic Studies, University of Stockholm, Sweden. I wish to acknowledge helpful comments on an earlier draft by the referee. He is absolved from responsibility for any remaining errors.

rest of the world may, however, have trade duties and investment taxes of its own. In that case the international terms of trade or lending facing the home country are different from those prevailing in the internal markets of foreign countries.

We wish to maximize a concave social utility function

$$(1) \quad U = U(C_1, C_2) = U(Y_1 - X_1, Y_2 - X_2)$$

where U_i will denote the partial derivative, C_1 and C_2 are consumption of good 1 and good 2, Y_1 and Y_2 are their outputs, and X_1 and X_2 are their exports.

The production constraint is written in the implicit form and takes account of the amount of capital invested abroad, and therefore not available for domestic use.

$$(2) \qquad \phi(Y_1, Y_2, F) = 0$$

where ϕ_i denotes the partial derivative. The production possibility curve is concave to the origin in the (Y_1, Y_2) plane shifts downward whenever F, the amount of foreign investment, increases. The balance of trade is in equilibrium both before and after a transfer of capital and takes account of both commodities and invisibles. Following the other writers above, we do not concern ourselves with the adjustment process during the transfer. Therefore:

$$(3) \qquad X_1 + X_2\pi + F\rho = 0$$

X_i positive represents an export and X_i negative, an import; π is the international terms of trade, with good 1 as numeraire; ρ is the foreign rate of interest, and F is the stock of external investment, in terms of good 1.

The free international mobility of capital leads to some complications with respect to the production possibility curve and the likelihood of incomplete specialization simultaneously for the home country and the rest of the world. We wish to make the production functions first-degree homogeneous, and at the same time to have incomplete specialization everywhere a reasonable possibility. With only two factors of production, Jones (1967) showed that the probability of incomplete specialization both at home and abroad was zero.[1] In order to remedy this difficulty we as-

[1] With only two factors of production and constant returns to scale, complete specialization can occur rather easily under free mobility of either factor. This has been shown by Jones, *op.cit* (1967), pp. 31–38. When technologies are the same everywhere, the allocation of capital between countries and the outputs of goods in each are indeterminate. All we know from the Lerner-Samuelson factor-price theorem is that real wages and the rate of interest are equalized under incomplete specialization everywhere, whether or not capital is mobile. When conditions of production are *not* the same everywhere, the one-to-one mapping between goods prices and factor rewards is inconsistent with the free mobility of capital and production of *both* goods in each country. For example, let one country have a technology such that the given price implies a lower reward to capital than abroad: mobility of capital prevents the downward adjustment of the interest rate; unit cost is too high in the capital-intensive sector and so capital flows out until that sector disappears. The wage will now be determined in the one remaining sector by price *and* the supply of capital. The latter in turn is adjusted to the prevailing rate of interest, while the wage may be either higher or lower than abroad, depending on the technology.

Parenthetically, J. S. Chipman has, in a forthcoming paper, analyzed a special two-by-two case where incomplete specialization of both countries *can* happen, but where terms of trade and factor prices are invariant so long as both goods are produced in both countries. That is, the world production-possibility curve is flat on the segment corresponding to incomplete specialization. The conditions are that in one good the production functions be the same, between countries, while in the other the labor coefficients differ; in addition (this is not stated explicitly by Chipman) capital coefficients must differ in the opposite sense, so that unit cost with equal factor prices can be the same in both countries. These highly special conditions illustrate how difficult it is to obtain incomplete specialization in the two-by-two case. Moreover, the invariance of price and factor rentals makes the case uninteresting for the analysis of optimal interference with trade and lending. I am indebted to Chipman for letting me see his paper.

When a third factor, land, is present the unspecialized system is generally compatible with a fixed capital rental. In terms of Samuelson's equations and unknowns, we have six marginal productivity equations and two factor-supply equations to determine four independent factor ratios, two independent factor allocations, and the rewards to labor and land (6+2=4+2+2).

sume three factors of production; labor, capital, and land. It is then possible to have interferences with trade and foreign investment in addition to differences in technology, together with perfect international mobility of capital (or for that matter, labor), and still have both goods being produced in both parts of the world.

We must, however, pay a price for introducing a third factor of production, in terms of other complications. For one thing, the presence of land in both production functions means that labor's and capital's marginal productivities depend on two factor ratios rather than just one. In order still to say something definite about factor rentals, when duties on trade or taxes on foreign investment are adjusted, we shall assume that one good is strongly labor-intensive, and the other strongly capital-intensive. We take this to mean that a rise in price and output of the labor intensive good always makes labor scarcer and raises its wage; conversely, the rental on capital always goes up when the output of the capital-intensive good increases. A second complication is that, with one-product firms, the presence of the third factor may cause the market-guided goods-transformation locus to take peculiar shapes. We therefore assume explicitly that this path is everywhere concave to the origin.[2]

Taking account of the production and balance-of-trade constraints, we now replace the function U with the function W, where we use the trade constraint to eliminate X_1, and we introduce the production constraint with a Lagrange multiplier.

$$W = U(Y_1 + X_2\pi + F\rho;\ Y_2 - X_2) - \lambda\phi(Y_1, Y_2, F) = W(X_2, F) \tag{4}$$

We wish to maximize the function W with respect to both the volume of good 2 traded and the volume of capital invested abroad. Domestic outputs, Y_1 and Y_2, and the international terms of trade and lending, π and ρ, are treated as dependent on both exports X_2, and stock of foreign investment F.

A full optimum is obtained when X_2 and F can each be freely varied by policy makers, in which case the two are independent for purposes of differentiation. After finding full optima with respect to X_2 and F in the two following Sections I and II, we shall deal with the partial-optimum problem of Jones in Sections III and IV. In finding the optimal volume of trade, we treat F as dependent assuming there is some policy constraint on changing the degree of restriction on foreign investment. Similarly, in finding the optimal magnitude of F, we assume in Section IV that restrictions on trade are not freely adjustable, so that X_2 is dependent on F. Section V deals with a country having an imperfect domestic labor market. This situation we take to be representative of many developing countries.

I. Optimal Trade Duty

The objective of this section is to find the optimal tariff when policy makers have complete freedom to adjust restrictions on foreign investment, and in fact do maximize the social-utility function with respect to both variables. Holding F constant, we differentiate (4) with respect to X_2 and set the derivative equal to zero.

[2] When there are three factors and two goods, and any firm produces only one good, one can construct examples in which a factor's marginal product rises in the expanding industry and falls in the contracting one. This occurs because a factor's marginal product depends on its ratios to *both* other factors. Taking that factor's wage as numeraire, this means that marginal cost is falling in the expanding sector and rising in the other. Hence opportunity cost changes in the wrong way for concavity. Just as important, there is no longer a one to one mapping between goods prices and factor prices. For a fuller discussion see Gehrels. The claim there, however, was not quite correct: it *is* possible for the transformation curve to be convex everywhere to the origin.

$$\begin{aligned} \partial W/\partial X_2 = {} & U_1(\partial Y_1/\partial X_2 + \pi \\ & + X_2 \partial\pi/\partial X_2 + F\partial\rho/\partial X_2) \\ & + U_2(\partial Y_2/\partial X_2 - 1) \\ & - \lambda(\phi_1 \partial Y_1/\partial X_2 \\ & + \phi_2 \partial Y_2/\partial X_2) = 0 \end{aligned} \tag{5}$$

Domestic outputs change whenever trade changes; and $\partial Y_2/\partial X_2$ is always positive, whether good 2 is an export or an import, implying that $\partial Y_1/\partial X_2$ must be negative. This follows from the condition $\phi_1 dY_1 + \phi_2 dY_2 = 0$, which is necessary for (2) to hold, given F constant, and that ϕ_1 and ϕ_2 are of the same sign. Thus dY_1 and dY_2 must be of opposite sign. The terms of lending depend on trade changes through adjustments in foreign production, given that the rest of the world is incompletely specialized. Which way ρ moves depends on which good is capital intensive.[3]

Because of perfect competition and my assumption of equality between social and private rates of substitution,

$$U_1 \partial Y_1/\partial X_2 + U_2 \partial Y_2/\partial X_2 = 0, \tag{6}$$

and

$$U_2/U_1 = p, \tag{7}$$

where p is the domestic price ratio in terms of good 1.

Using (6) and (7), (5) now reduces to

$$\begin{aligned} & \pi + X_2 \partial\pi/\partial X_2 + F\frac{\partial\rho}{\partial X_2} - p \\ & - \frac{\lambda}{U_1}\left(\frac{\phi_1}{\phi_2} + \frac{\partial Y_2}{\partial X_2} \Big/ \frac{\partial Y_1}{\partial X_2}\right)\frac{\phi_2}{\partial Y_1/\partial X_2} \end{aligned} \tag{8}$$

Since ϕ is the implicit form of the transformation function, ϕ_1/ϕ_2 is the marginal rate of transformation between the two goods, i.e., the amount of good 2 obtained by giving up a unit of good 1 in production. But the expression in parentheses is zero, so that the ratio of production changes brought about by the change in trade is equal to the marginal rate of domestic transformation. The optimal tariff rate can now be written as

$$\begin{aligned} t = {} & \frac{\pi - p}{\pi} \\ & - \frac{X_2}{\pi}\frac{\partial\pi}{\partial X_2} - \frac{F}{\pi}\frac{\partial\rho}{\partial X_2} \\ = {} & \frac{1}{\eta}\left[1 + \frac{F\rho}{X_2\pi}\frac{E_\rho}{E_\pi}\right] \end{aligned} \tag{9}$$

which result agrees with those of Kemp and Jones.[4] The coefficient η is the foreign price elasticity of demand for home-country exports of good 2 (or foreign supply elasticity, when good 2 is an import of the home country); $F\rho/X_2\pi$ is the ratio of investment income to value of good 2 traded; and E_ρ/E_π is the elasticity of the interest rate with respect to the terms of trade. Through the "magnification effect" (see Jones 1965) this is generally greater than one. The reason for this is that a price is a weighted average of factor-rental changes; as output shifts, wages and the interest rate move in opposite directions, owing to our strong-intensity assumption above. On the other hand, $F\rho/X_2\pi$ is generally less than one. The second term in the bracket could therefore be greater or less than one in absolute value, and have either sign. Its effect on the size of the

[3] More generally, with constant returns to scale, but variable returns to proportion, the factor-price theorem of Samuelson states that (a) with an equal number of goods and factors there is a one-to-one mapping between goods prices and factor rentals, so that the terms of trade and lending must *both* move if either of them moves; (b) with more factors than goods (as is the case here with three factors and two goods, and a fortiori with complete specialization) there is a unique mapping from goods prices and factor supplies to factor rentals. Thus, with given factor supplies there is again a connection between prices and rentals, and the latter must change with the former.

[4] The equivalent of expression (9) has been found both by Kemp and by Jones. See Jones, (1967) equations (9) and (11) and pp. 8–16; also Kemp, equation (3) and pp. 792–95.

optimal duty (which could become a subsidy) therefore depends on particular assumptions.

To examine a particular case, let us assume that the home country is a net exporter of capital, and that its export good—let it be good 2—is capital intensive both at home and abroad. An increase in t, its trade duty, then increases π; it also reduces the volume of trade, causes an increase in foreign production of good 2, and leads to a shift of home production away from good 2. Which way do the terms of lending move? From our assumptions they must move unambiguously in favor of the home country, because under full optimization, F the stock of foreign capital owned by the home country, is independent of the tariff rate. Thus the only fact of significance is that the rest of the world uses its capital more intensively (i.e., with more labor) in both employments than it did before.

One can easily see that a debtor country which also imports the capital-intensive good improves its terms of borrowing along with its terms of trade, when it raises its trade duty. This rather credible pair of cases suggests that the consequence of the terms-of-lending consideration for both creditor and debtor nations is the same: Each individually has an interest in restricting trade by more than the degree indicated by trade elasticity alone. Even if foreign demand elasticity were infinite, there may still be an argument for restricting trade—providing, of course, that there is no retaliation.

Without additional discussion it should be clear that when the home country is a capital exporter but exports the labor-intensive good, consideration of the return on foreign investment reduces the optimal duty on trade. Thus one can not make a general case either for higher or for lower duties on trade (before foreign retaliation) as a result of considering effects on the terms of borrowing in addition to effects on the terms of trade.

Finally, it is natural to ask whether the situation is changed markedly when, contrary to the present assumption, the rest of the world is specialized completely. Our differentiating procedure was to hold F constant while differentiating partially with respect to X_2. If the outside world is specialized in the numeraire good, then with a constant employment of all factors in that one good, the nominal return to capital abroad is independent of the terms of trade, and so of the duty. Hence the optimal tariff in (9) reduces to the reciprocal of price elasticity. But if the outside world is specialized in the non-numeraire good, then the rate of return on capital does depend on price. That is, the marginal physical productivity of capital stays constant as before, but its value productivity must change in the same proportion as price. (The elasticity of capital's return with respect to price is plus or minus unity.)[5]

We conclude this section by pointing out that it is possible to extend the procedure above without difficulty to that of n-goods produced and traded. Leaving its derivation to Appendix A, we obtain

$$(10) \qquad t_k = \sum_{i=2}^{n} \frac{V_i}{V_k} \cdot \frac{1}{\eta_{ki}} - \frac{F\rho}{V_k} \cdot \frac{E_\rho}{E_{X_k}}$$

where η_{ki} is the foreign demand elasticity of the kth good traded with respect to the ith price, while V_i/V_k is the ratio of values in trade, of the ith and kth goods. The first term is therefore the weighted sum of reciprocals of price elasticities of the kth good with respect to all $n-1$ prices. The second term is the elasticity of the foreign

[5] Kemp, Section III, does not distinguish between the two cases, even though he holds the foreign capital stock fixed. It is, however, unclear if he is referring to physical capital or value of capital in terms of the numeraire. Jones has the first half of the conclusion in the text above, but not the second half. See Jones (1967), Section III.

rate of interest with respect to the quantity of the kth good traded and weighted by the ratio of interest income to the value of the kth good in trade.

II. Optimal Investment Tax

In order to find the utility-maximizing tax on foreign investment, we now hold X_2 fixed and differentiate partially with respect to F. Domestic output is a function of F because a unit increase of capital invested abroad means a unit reduction of capital at home. This generally affects output of both goods, but differently according to their relative factor intensities.[6] The terms of trade are also generally a function of F because goods prices are usually not independent of factor prices, and when the latter are affected by changes in F, the former are also. We obtain

$$\begin{aligned}\frac{\partial W}{\partial F} = U_1\left(\frac{\partial Y_1}{\partial F} + X_2\frac{\partial \pi}{\partial F} + \rho + F\frac{\partial \rho}{\partial F}\right)\\ + U_2\left(\frac{\partial Y_2}{\partial F}\right) - \lambda\left(\phi_1\frac{\partial Y_1}{\partial F}\right.\\ \left.+ \phi_2\frac{\partial Y_2}{\partial F} + \phi_F\right) = 0\end{aligned} \tag{11}$$

From the side condition we obtain directly that

$$\begin{aligned}-r = -\frac{\phi_F}{\phi_1} = \frac{\partial Y_1}{\partial F} + \frac{\phi_2}{\phi_1}\frac{\partial Y_2}{\partial F}\\ = \frac{\partial Y_1}{\partial F} + p\frac{\partial Y_2}{\partial F}\end{aligned} \tag{12}$$

This states that the rate of interest, r, is equal to the marginal contribution of capital to national product, measured in units of the numeraire good. This, in turn, is the negative of the sum of changes in output due to the transfer of a unit of capital abroad. Using, in addition, that $U_2/U_1 = p$, we reduce (11) to

$$\begin{aligned}X_2\frac{\partial \pi}{\partial F} + \rho + F\frac{\partial \rho}{\partial F}\\ + p\frac{\partial Y_2}{\partial F} + \frac{\partial Y_1}{\partial F} = 0,\end{aligned} \tag{13}$$

and then to

$$\begin{aligned}\frac{\rho - r}{\rho} = -\frac{X_2}{\rho}\frac{\partial \pi}{\partial F} - \frac{F}{\rho}\frac{\partial \rho}{\partial F}\\ = -\left(\frac{X_2}{\rho}\frac{\pi}{F}\frac{E_\pi}{E_\rho} + 1\right)\frac{E_\rho}{E_F}\end{aligned} \tag{14}$$

where E_π/E_ρ is the elasticity of foreign price with respect to the rate of interest. This is less than one, because price change is the weighted average of factor-rental changes, and, under our assumptions above, the wage rate falls when the interest rate rises. The ratio E_ρ/E_F is the elasticity of the foreign interest rate with respect to foreign investment by the home country. Let us again assume that good 2, the capital-intensive good, is exported by the lending country.

Some idea of the order of magnitude of the optimal tax on investment can be obtained by noting, as did MacDougall, that the elasticity of demand of a borrowing country for capital from abroad is the share of foreign-owned capital in its capital stock, times its total elasticity of demand for capital. The first term in the parentheses (14) can be either weaker or stronger than the second, because $X\pi/F\rho > 1$, being the ratio of visible to invisible exports in value terms, and $E_\pi/E_\rho < 1$. If their product were unity, the tax rate would therefore be about twice the elasticity of the foreign rate with respect to foreign investment.

[6] Under the assumptions made for the factor-price equalization theorem, and in the two-good two-factor case, the Rybczynski theorem states that exporting a unit of capital will cause one output to fall and the other to increase. See T. Rybczynski. We can not invoke the theorem here because we have three, rather than two factors.

In the converse case, where the home country is capital-poor and thus both a net debtor and an importer of good 2, the optimal restriction of foreign borrowing is made greater by the fact that the foreign price of imported good 2 is reduced by the same action. In (9) X_2 and F from their definitions are both negative, while $\partial\pi/\partial F$ and $\partial\rho/\partial F$ remain negative; hence $\rho - r$ is now *negative*. The foreign return on capital is held below the return in the borrowing country.

We conclude that in the cases here described, the terms-of-trade and terms-of-lending effects reinforce each other for *both* creditor and debtor countries. Barring retaliation, optimal restrictions on trade and on lending will be more severe than when cross effects are neglected. It seems unnecessary to explore the case of X_2 being labor-intensive; but one can easily follow the same line of reasoning to show that the optimal restrictions on trade and lending would both be reduced in this case, and this is true for both creditor and debtor countries.

This result is related to that of Kemp and of Jones, if different in form. The reason for the difference lies in their maximizing social utility with respect to price, and to quantity of investment.[7] Intuitively, what they have done can be viewed as follows: In the neighborhood of the optimum, transferring one unit of capital causes market disequilibria at initial prices both at home and abroad. Since foreign price can not be adjusted, equilibrium is restored by adjusting home-country prices, and so the tariff. By contrast, we have made quantity of exports independent, and allowed the secondary adjustment to the transfer of capital to take place in prices everywhere. Here too, some secondary adjustment of the tariff is needed.

For the n-goods case, leaving the derivation to Appendix B, we obtain the expression

$$(15) \qquad \rho - r = - \sum_{i}^{n} X_i \frac{\partial \pi_i}{\partial F} - F \frac{\partial \rho}{\partial F}$$

The excess of the foreign over the domestic rate of interest is thus the weighted response of the foreign interest rate, adjusted for all the price responses to the redistribution of the capital stock between countries. If, on balance, the home country exports capital-intensive goods, which are also capital-intensive abroad, then $\sum X_i \partial\pi_i/\partial F$ is negative, as is $\partial\rho/\partial F$. The previous conclusion, that optimal restrictions on lending are made more severe, is substantially unchanged.

III. Partial Optimization of Trade Duty

Governments are not always able to vary both the trade duty and the investment tax in order to achieve the kind of optimum defined above. It is therefore of interest to examine more generally the question considered by Jones, namely, how an optimal duty is obtained when the tax on investment cannot be varied.[8] This section will show that the optimal duty subject to a given foreign-investment tax is higher than under full optimization, when the export good is also capital intensive. The corresponding modification of the optimal investment tax is left to the following section.

The main difference in finding an optimum lies in the fact that F, the amount of

[7] This leads Jones to conclude that there can be a non-zero optimal tax on foreign investment only when there is a difference between domestic and foreign prices of goods, since under his assumptions neither price nor the interest rate in the unspecialized country is dependent on the stock of capital received from abroad. See Jones (1967) p. 12, and equation (12) and Kemp, p. 793, equation (2c), and p. 795, (5) and (6).

[8] Jones treats only the special case where the tax on investment is held at zero, so that $r = \rho$. See Jones (1967), pp. 23–31.

$$\begin{aligned}\frac{\partial W}{\partial X_2} = {} & U_1\left(\frac{\partial Y_1}{\partial X_2} + \frac{\partial Y_1}{\partial F}\frac{\partial F}{\partial X_2} + \pi + X_2\frac{\partial \pi}{\partial X_2} + F\frac{\partial \rho}{\partial X_2} + \rho\frac{\partial F}{\partial X_2}\right) \\ & + U_2\left(\frac{\partial Y_2}{\partial X_2} + \frac{\partial Y_2}{\partial F}\frac{\partial F}{\partial X_2} - 1\right) \\ & - \lambda\left[\phi_1\left(\frac{\partial Y_1}{\partial X_2} + \frac{\partial Y_1}{\partial F}\frac{\partial F}{\partial X_2}\right) + \phi_2\left(\frac{\partial Y_2}{\partial X_2} + \frac{\partial Y_2}{\partial F}\frac{\partial F}{\partial X_2}\right) + \phi_F\frac{\partial F}{\partial X_2}\right]\end{aligned} \tag{16}$$

foreign investment, is no longer independent of X_2, the amount of exports. Differentiating (4) with respect to X_2 gives equation (16).

In order to reduce this expression, we have the following three relations:

$$\frac{\partial Y_1}{\partial X_2} + p\frac{\partial Y_2}{\partial X_2} = 0 \tag{17}$$

$$\frac{\partial Y_1}{\partial X_2} + \frac{\phi_2}{\phi_1}\frac{\partial Y_2}{\partial X_2} = 0 \tag{18}$$

$$\frac{\partial Y_1}{\partial F}\frac{\partial F}{\partial X_2} + p\frac{\partial Y_2}{\partial F}\frac{\partial F}{\partial X_2} = -r\frac{\partial F}{\partial X_2} \tag{19}$$

Since all three have been used in reductions above, they do not need further discussion. Equation (16) now becomes

$$\begin{aligned}\frac{\pi - p}{\pi} = {} & -\frac{(\rho - r)}{\pi}\frac{\partial F}{\partial X_2} \\ & - \frac{X_2}{\pi}\frac{\partial \pi}{\partial X_2} - \frac{F}{\pi}\frac{\partial \rho}{\partial X_2}\end{aligned} \tag{20}$$

This expression differs in form from its counterpart (9) above only in taking account of any adjustment of the quantity of foreign investments to trade. In the special case where $\rho = r$ (the only one considered by Jones 1967), the expression reduces back to (9). However, this is not to say that the numerical result for the optimal duty is the same, because the values of the two remaining terms will differ according to the degree of interference with investment.

Given again that X_2 is exported by the home country and is capital intensive, and that there is incomplete specialization at home and abroad, an increase of the duty encourages foreign investment because foreign output of X_2 increases. Thus $\partial F/\partial X_2 < 0$, and $\partial \pi/\partial X_2 < 0$ just as before. If the foreign rental on capital is higher than that at home the gain from restricting trade is greater than in the independent case, and the optimal duty is higher.

Suppose now that the home country is instead the debtor, and that X_2 is the capital-intensive *import* good. Does the same expression show that a debtor country would restrict trade more severely than when it considers only the terms of trade and not the terms of lending? The answer is clearly yes because the signs of the three derivatives are the same, but $\rho < r$, $F < 0$, $X_2 < 0$, and $\pi < p$.

An intuitive interpretation of the extra term $\frac{(\rho - r)}{\pi}\frac{F}{X_2}$ is that when a lending country sends additional capital abroad because of a change in trade, and the foreign return is higher than the return to home, each unit gives it a net gain of $(\rho - r)$. This is additional to any induced change in the terms of lending. Conversely, when a debtor country obtains more foreign capital because of a change in trade, the gain for it is $(r - \rho)$ per unit, where this measures the excess of the domestic return over the foreign cost per unit of capital.

$$\frac{\partial W}{\partial F} = U_1\left(\frac{\partial Y_1}{\partial F} + \frac{\partial Y_1}{\partial X_2}\frac{\partial X_2}{\partial F} + X_2\frac{\partial \pi}{\partial F} + \pi\frac{\partial X_2}{\partial F} + \rho + F\frac{\partial \rho}{\partial F}\right)$$

$$(21)\qquad + U_2\left(\frac{\partial Y_2}{\partial F} + \frac{\partial Y_2}{\partial X_2}\frac{\partial X_2}{\partial F} - \frac{\partial X_2}{\partial F}\right)$$

$$-\lambda\left[\phi_1\left(\frac{\partial Y_1}{\partial F} + \frac{\partial Y_1}{\partial X_2}\frac{\partial X_2}{\partial F}\right) + \phi_2\left(\frac{\partial Y_2}{\partial F} + \frac{\partial Y_2}{\partial X_2}\frac{\partial X_2}{\partial F}\right) + \phi_r\right] = 0$$

IV. Partial Optimization of Investment Tax

We now assume that the tariff is fixed and that the tax on foreign investment is the policy variable. This problem was ingeniously analyzed by Jones but only for the particular case of free commodity trade (where $\pi = p$).[9]

We differentiate (4) with respect to F, to obtain equation (21). This time the side condition gives us that

$$(22)\qquad r = \frac{\phi_r}{\phi_1} = -\left(\frac{\partial Y_1}{\partial F} + \frac{\partial Y_1}{\partial X_2}\frac{\partial X_2}{\partial F}\right) - \frac{\phi_2}{\phi_1}\left(\frac{\partial Y_2}{\partial F} + \frac{\partial Y_2}{\partial X_2}\frac{\partial X_2}{\partial F}\right)$$

This again states that the loss of domestic output due to the transfer of a unit of capital matches the domestic rate of interest. After eliminations we get from (21) that

$$(23)\qquad \rho - r = -X_2\frac{\partial \pi}{\partial F} - F\frac{\partial \rho}{\partial F} - (\pi - p)\frac{\partial X_2}{\partial F}$$

Let us first suppose that the home country is a creditor and an exporter of the capital-intensive good, X_2. It is easily verified that

$$\frac{\partial X_2}{\partial F} < 0, \qquad \frac{\partial \rho}{\partial F} < 0, \qquad \frac{\partial \pi}{\partial F} < 0,$$

and $(\pi - p) > 0$ when there is a positive duty on trade. In expression (23), $(\rho - r)$ must therefore be positive because every term on the right is positive. The fact that exports of the capital-intensive good react negatively to capital exports, and thereby cause a loss per unit of export reduction measured by $(\pi - p)$, makes it worthwhile to restrict foreign investment by more than when trade is independent.

The expression works in an exactly symmetric way when the home country is a debtor and an importer of capital-intensive goods, because the derivatives are unchanged in sign, and $(\pi - p)$, F, and X_2 are now all negative. Restriction of foreign borrowing has an additional gain due to the increased importation of X_2, weighted by $(p - \pi)$ per unit.

We conclude that, given the values of the price response and interest reponse terms in the optimal tax on investment equation (23), the optimal restriction on the international flow of capital is increased by the response of trade volume to investment. This conclusion however depended on the exportable good 2 being capital intensive. If good 2 were instead labor intensive the sign of the export response to foreign investment becomes positive ($X_2/F > 0$) and the extra term containing the price difference has the effect of reducing the optimal tax.

V. Optimal Interference for a Developing Country

The capital-poor country may in addi-

[9] Jones (1967) pp. 15–23.

tion have distortions in the domestic market for factors of production, with the consequences for allocation discussed by Hagen, Bhagwati and Ramaswami, and Fishlow and David. Let the Y_2 sector (which is capital-intensive and import-competing) face a money wage rate higher than its opportunity cost in terms of labor's marginal product in the Y_1 sector. The consequence of this is (a) that Y_2's money cost overstates its opportunity cost, so that its relative share in output is reduced from the perfectly competitive share; and (b), that the production-possibility curve is shifted inward because Y_1 uses too small a proportion of labor to other factors and Y_2, too high a proportion. Marginal rates of transformation between factors are not equal between sectors. Allocation is inefficient on two counts and can be corrected fully only by correcting imperfections in the labor market. We wish to see how the rules above are modified when a labor-market distortion is introduced. That is, we seek a second-best optimum, subject to a given constraint in the labor market. The utility function to be maximized is the same, but the constraints are different. We have, as before,

$$(24)\qquad \begin{aligned} U &= U(C_1, C_2) \\ &= U(Y_1 - X_1, Y_2 - X_2) \end{aligned}$$

The production constraint is now written as

$$(25)\qquad \psi(Y_1, Y_2, F) = 0,$$

where, for a given domestic capital stock, the production-possibility curve lies inside that for ϕ (equation (2)) everywhere but at the end points, because of the labor-market imperfection. We have, in addition, that

$$(26)\qquad U_2/U_1 = p = -b\frac{\partial Y_1}{\partial Y_2} = b\frac{\psi_2}{\psi_1},\ b > 1$$

This states that the marginal utility of good 2 in terms of good 1 is greater than its opportunity cost in the constant ratio b. The necessary condition for a full optimum with respect to the duty becomes

$$(27)\qquad \begin{aligned} \frac{\partial W}{\partial X_2} = U_1\left(\frac{\partial Y_1}{\partial X_2} + \pi + X_2\frac{\partial \pi}{\partial X_2}\right. & \\ \left. + F\frac{\partial \rho}{\partial X_2}\right) + U_2\left(\frac{\partial Y_2}{\partial X_2} - 1\right) & \\ - \left(\frac{\partial \psi}{\partial Y_1}\frac{\partial Y_1}{\partial Y_2} + \frac{\partial \psi}{\partial Y_2}\frac{\partial Y_2}{\partial X_2}\right) &= 0 \end{aligned}$$

Using (26) this reduces to

$$(28)\qquad \begin{aligned} U_1\Big[(1-b)\frac{\partial Y_1}{\partial X_2} + \pi + X_2\frac{\partial \pi}{\partial X_2} & \\ + F\frac{\partial \rho}{\partial X_2}\Big] - U_2 &= 0 \end{aligned}$$

$$(29)\qquad \begin{aligned} \frac{p-\pi}{\pi} = \frac{X_2}{\pi}\frac{\partial \pi}{\partial X_2} + \frac{F}{\pi}\frac{\partial \rho}{\partial X_2} & \\ - \frac{(b-1)}{\pi}\frac{\partial Y_1}{\partial X_2} & \end{aligned}$$

In examining this expression, note that now $X_2<0$ because X_2 is the import; $\partial\pi/\partial X_2<0$; $F<0$ because the country is a net debtor; $\partial\rho/\partial X_2<0$ because good 2 is capital-intensive, and reducing its importation (dX_2 positive) causes the foreign interest rate to fall. Further, $\partial Y_1/\partial X_2<0$, and $b-1>0$. Consequently, all three terms on the right-hand side are positive, and the effect of the labor-market imperfection is *to increase* the optimal rate of duty on imports. In common-sense terms, the overstated cost of good 2, and its consequent underproduction, is mitigated by the additional protection given to it. Even if the home country had no influence on the terms of trade or lending, it would still be worthwhile to restrict imports.

In order to find the optimal restriction on foreign borrowing, we hold X_2 constant and differentiate (24) with respect to F.

$$\frac{\partial W}{\partial F} = U_1\left(\frac{\partial Y_1}{\partial F} + X_2\frac{\partial \Pi}{\partial F} + \rho + F\frac{\partial \rho}{\partial F}\right) + U_2\frac{\partial Y_2}{\partial F} - \lambda\left(\psi_1\frac{\partial Y_1}{\partial F} + \psi_2\frac{\partial Y_2}{\partial F} + \psi_F\right) \tag{30}$$

As before

$$-r = -\psi_F/\psi_1 = \frac{\partial Y_1}{\partial F} + \frac{\psi_2}{\psi_1}\frac{\partial Y_2}{\partial F} = \frac{\partial Y_1}{\partial F} + \frac{p}{b}\frac{\partial Y_2}{\partial F} \tag{31}$$

$$r - p = X_2\frac{\partial \Pi}{\partial F} + F\frac{\partial \rho}{\partial F} + p\frac{(b-1)}{b}\frac{\partial Y_2}{\partial F} \tag{32}$$

Because good 2 is capital-intensive, and importing more capital tends to reduce foreign production of good 2, and therefore to raise its foreign price, we have $\partial\pi/\partial F < 0$; importing more capital encourages home production of good 2, so that $\partial Y_2/\partial F < 0$; and $\partial\rho/\partial F < 0$ because importing more capital tends to raise the foreign rate of interest. Remembering that X_2 and F are both negative, the terms of trade effect, if any, reinforces the terms of lending effect. But of greater interest is that the domestic cost distortion works counter to the first two effects. The first two terms are positive, but the third term is negative. The common sense of this is that the overstated costs of the import-competing sector, owing to excessive money wages, can be mitigated by reducing the home interest rate relatively to the international rate. If the home country has no influence on the international terms of trade or lending, expression (32) states that a developing country should *subsidize* foreign borrowing. That is, it should make the domestic rate of interest lower than the foreign rate.

VI. Summary

The general conclusion of this paper is that the optimal restriction on trade or foreign investment is changed by the interrelation between goods prices and factor prices. Whether each restriction is increased or decreased thereby depends on the factor-intensities of the traded goods; but in general the two kinds of interference would be altered in the same direction.

When there is complete freedom to adjust both the duty rate and the tax on foreign investment, the optimal duty is increased for the exporter of capital-intensive goods if that country is also a creditor; the same statement is true for a country which exports labor-intensive goods and is a debtor. The opposite conclusions follow if factor intensities are switched.

The optimal tax on foreign investment or borrowing is increased under exactly the same conditions.

When the difference between international and domestic rate of return on capital is taken as fixed, the optimal duty is further modified because of induced transfers of capital. For the capital-intensive exporter and lender, the induced outflow of capital brings additional gain. For the labor-intensive exporter and borrower, the induced inflow brings an added advantage and implies a further increase of the duty.

When the international and domestic price ratios differ by a fixed proportion, the induced increase of trade makes it profitable for both creditor and debtor country to impose a larger tax on lending or borrowing than under 2.

When the debtor country also suffers from overstated costs in the import-competing, capital-intensive sector, because of differential wages, it becomes worthwhile to raise the trade duty further but to *reduce* the tax on foreign-owned capital.

Appendix

We shall here derive expressions (10) and

(15) in the text, which give the optimal tariff and optimal investment tax for the case of n goods.

A. Optimal Tariff

Expressions (1) through (4) of the text were the collective utility function, the production function in implicit form, the balance-of-payments constraint, and the constrained utility function. They now become

$$(1') \quad U = U(C_1, C_2, \ldots, C_n) = U(Y_1 - X_1, Y_2 - X_2, \ldots, Y_n - X_n)$$

$$(2') \quad \phi(Y_1, Y_2, \ldots, Y_n, F) = 0$$

$$(3') \quad X_1 + \sum_2^n X_i\pi_i + F\rho = 0$$

$$(4') \quad W = U\left(Y_1 + \sum_2^n X_i\pi_i + F\rho; Y_2 - X_2, Y_3 - X_3, \ldots, Y_n - X_n\right) - \lambda\phi(Y_1, Y_2, \ldots, Y_n, F) = 0$$

Just as before they describe the social utility function, the production function, and the balance-of-trade constraint. All but one of the X_i's are independent. We have chosen to make X_1 a function of the other $n-1$ X_k's, and have replaced it by means of the trade constraint (12). As in the two-goods case, all outputs are functions of each X_k, as are the $n-1$ international prices π_i, and the rate of return on foreign investment, ρ.

$$(5') \quad \frac{\partial W}{\partial X_k} = U_1\left(\frac{\partial Y_1}{\partial X_k} + \pi_k + \sum_2^n X_i \frac{\partial \pi_i}{\partial X_k} + F\frac{\partial \rho}{\partial X_k}\right) + U_{k-1}\frac{\partial Y_{k-1}}{\partial X_k} + U_k\left(\frac{\partial Y_k}{\partial X_k} - 1\right) + U_n\frac{\partial Y_n}{\partial X_k} - \lambda\left(\phi_1\frac{Y_1}{X_k} + \phi_k\frac{Y_k}{X_k} + \phi_n\frac{Y_n}{X_k}\right) = 0$$

$$k = 2, 3, \ldots, n$$

In the same way as before, we have

$$(6') \quad \sum_2^n U_i\frac{\partial Y_i}{\partial X_k} + U_1\frac{\partial Y_1}{\partial X_k} = 0,$$

and

$$(7') \quad U_i/U_1 = p_i = -\partial Y_1/\partial Y_i \qquad i = 2, 3, \ldots, n$$

With the use of (6′) and (7′), (5′) reduces to

$$(8') \quad \frac{\partial W}{\partial X_k} = \pi_k + \sum_2^n X_i\frac{\partial \pi_i}{\partial X_k} + F\frac{\partial \rho}{\partial X_k} - p_k = 0$$

The expression for the optimal tariff is therefore

$$t_k = \frac{\pi_k - p_k}{\pi_k} = -\sum_2^n \frac{X_i}{X_k}\frac{\pi_i}{\pi_k}\frac{\partial \pi_i}{\partial X_k}\frac{X_k}{\pi_i} - \frac{F\rho}{X_k\pi_k}\frac{E_\rho}{E_{X_k}}$$

Stated more simply, this is

$$(10') \quad t_k = \sum_2^n \frac{V_i}{V_k}\frac{1}{\pi_{ki}} - \frac{F\rho}{V_k}\frac{E_\rho}{E_{X_k}}$$

B. Optimal Investment Tax

For the n-goods case, we differentiate (4′) with respect to F and obtain

$$(11') \quad \frac{\partial W}{\partial F} = U_1\left(\frac{\partial Y_1}{\partial F} + \sum_2^n X_i\frac{\partial \pi_i}{\partial F} + \rho + F\frac{\partial \rho}{\partial F}\right) + \sum_2^n U_i\frac{\partial Y_i}{\partial F} - \lambda\left(\phi_i\frac{\partial Y_1}{\partial F} + \sum_2^n \phi_i\frac{\partial Y_i}{\partial F} + \phi_F\right) = 0$$

We can simplify this by using the fact that

$$(12') \quad -r = -\frac{\phi_F}{\phi_1} = \frac{\partial Y_1}{\partial F} + \sum_2^n \frac{\phi_i}{\phi_1}\frac{\partial Y_1}{\partial F}$$

That is, the loss of home output due to a unit transfer of capital is the sum of all the output changes valued in terms of the numeraire good, and this in turn matches the rate of interest, under competition. Using (25) reduces the expression (24) to

$$(15)\quad \rho - r = -\sum_{2}^{n} X_i \frac{\partial \pi_i}{\partial F} - F\frac{\partial \rho}{\partial F}$$

REFERENCES

J. Bhagwati and V. K. Ramaswami, "Domestic Distortions, Tariffs, and The Theory of Optimum Subsidy," *J. Polit. Econ.*, Feb. 1963, *71*, 44–50.

J. S. Chipman, "International Trade with Capital Mobility: A Substitution Theorem," in J. N. Bhagwati et al, eds., *Trade, Balance of Payments, and Growth; Essays in Honor of Charles Kindleberger*, Cambridge, Mass. 1971.

A. Fishlow and P. David, "Optimal Resource Allocation in an Imperfect Market Setting," *J. Polit. Econ.*, Dec. 1961, *69*, 529–46.

F. Gehrels, "Factor Marginal Products and Decreasing Opportunity Cost," *Amer. Econ. Rev.*, Mar. 1965, *55*, 114–21.

J. de V. Graaff, "On Optimum Tariff Structures," *Rev. Econ. Stud.*, 1949, No. 1, *17*, 47–59.

E. Hagen, "An Economic Justification of Protectionism," *Quart. J. Econ.*, Nov. 1958, *72*, 496–514.

G. E. Jasay, "The Social Choice between Home and Overseas Investment," *Econ. J.*, Mar. 1960, *70*, 105–13.

R. W. Jones, "International Capital Movements and the Theory of Tariffs and Trade," *Quart. J. Econ.*, Feb. 1967, *81*, 1–38.

———, "The Structure of Simple General Equilibrium Models," *J. Polit. Econ.*, Dec. 1965, *73*, 557–72.

M. C. Kemp, "The Gain from International Trade and Investment: A Neo-Heckscher-Ohlin Approach," *Amer. Econ. Rev.*, Sept. 1966, *56*, 788–809.

G. D. A. MacDougall, "The Benefits and Costs of Private Investment from Abroad: A Theoretical Approach," *Econ. Rec.*, Mar. 1960, *36*, 13–35, reprinted in R. Caves and H. Johnson, eds., A.E.A. *Readings in International Economics*, Homewood 1968, 172–97.

T. Negishi, "Foreign Investment and the Long-Run National Advantage," *Econ. Rec.*, Dec. 1965, *41*, 628–31.

V. K. Ramaswami, "International Factor Movement and The National Advantage," *Economica*, Aug. 1968, *35*, 309–10.

T. Rybczynski, "Factor Endowments and Relative Commodity Prices," *Economica*, Nov. 1955, *22*, 336–41.

P. A. Samuelson, "Prices of Factors and Goods in General Equilibrium," *Rev. Econ. Stud.*, Oct. 1953, *21*, 1–20.

COMMUNICATIONS

Large Industrial Corporations and Asset Shares: Comment

By David R. Kamerschen*

In a recent issue of this *Review*, David Mermelstein, following the tradition of A. D. H. Kaplan and N. R. Collins and L. E. Preston, examined the changing shares of the 100 largest corporations, on a decade-by-decade basis over the period 1909–64. Mermelstein finds evidence of an increased ability in recent decades for the largest corporations to maintain their share of total assets. He then considers the "*progressive* or *growing* advantages that are possessed by these largest corporations" (p. 539). The purpose of this communication is to comment briefly on the alternative explanations offered by Mermelstein for the increased stability of the asset shares of the 100 largest corporations.

I think Mermelstein has made an important contribution to the empirical literature of industrial organization. To be sure, there are some serious data limitations in any study of the sort he has undertaken, but I do not feel this is an insuperable difficulty. Much more serious are the theoretical limitations underlying the "turnover" method. I do not think Mermelstein has satisfactorily answered the "hostile comments and reviews" (p. 531) that have been levied at the Kaplan turnover approach. Since most of these chastisements are cited in Mermelstein, I will not refer to them with the sole exception of quoting the following, and as far as I can see, yet unanswered criticism of George J. Stigler.

> The statistical universe of the hundred or two hundred largest corporations is inappropriate to studies of monopoly and competition, and we may hope that this [referring to Kaplan] will be the last study to fall prey to its dramatic irrelevance. For Kaplan's central idea—that the extent of instability in the relative fortunes of the leading firms is an informative symptom of competition—is important and deserves to be applied on a correct, industry basis. [1969, p. 338]

Although Mermelstein did not, of course, apply the Kaplan turnover method on "a correct industry basis," let me drop this general criticism in order to concentrate on his rationale for his empirical results. I have no serious questions concerning his first explanation, viz. the advantages of size.[1] Two minor objections that might be voiced are 1) his failure to cite what many would regard as probably the best empirical work that has been done on this topic by Marshall Hall and Leonard Weiss[2] and 2) his failure to tie in his argument more explicitly with the enormous rise in conglomeration in recent years (e.g., see Kamerschen 1970).

It is his second explanation that is my primary concern. This explanation is, he claims, a corollary of the basic Berle-Means-Larner hypothesis of the separation of ownership and management.[3] Mermelstein argues, fol-

* Professor of economics, University of Missouri. I am grateful to Richard L. Wallace for his valuable comments and suggestions. The research reported here was financially assisted by a University of Missouri Summer Research Fellowship.

[1] Stigler (1968) argues imperfections in the capital market is too often given as an explanation without any empirical evidence as to the transaction costs (see Kamerschen 1969b).

[2] Although a recent study by Richard Arnould cites both my work and Hall and Weiss as having demonstrated the "relationship between absolute size and the cost of capital to be highly significant and positive" (p. 74), this was a secondary theme in my paper and was not exploited with anywhere near the care and finesse with which Hall and Weiss operated.

[3] I refer to the hypothesis in this way since Berle and Means first suggested the thesis, and Larner recently documented that separation of ownership and management has in fact occurred in our economy over the last fifty years.

lowing R. Joseph Monsen and Anthony Downs, that if all firms were owner controlled there would be more variability in earnings and hence asset shares. "If instead all firms are managerial firms, then we would expect less variability in earnings and similarly fewer shifts in asset shares" (Mermelstein, p. 540).

In the end, whether owner and manager controlled firms perform differently is an empirical question. And it seems to me that the evidence is at best ambiguous with regard to the point Mermelstein is trying to make.

While Mermelstein does not cite any of the statistical findings on this point, at least four such studies have appeared. Three of these studies by Brian Hindley,[4] Robert Larner, and Kamerschen (1968) provide no evidence to support the position that manager controlled firms perform differently in terms of profits than do owner controlled firms. The findings in these studies are now included in at least one introductory textbook. Thus Robert Lipsey and Peter Steiner (p. 370) state:

> When Professor Larner sought to explain the significance of the difference between corporations with and without dominant ownership groups, he hypothesized that the manager-controlled companies should have lower profits and show less variation in profits if they were trying to avoid risks. The evidence when examined led him to reject the view of a significant difference in behavior.[5]

Monsen, John Chiu, and D. E. Cooley did obtain results contrary to those of Hindley, Larner, and Kamerschen. That is, Monsen et al., found that the control status in the firms did influence profitability. While there has been at least one attempt to reconcile these differences (Kamerschen, 1969a) no definite conclusion can yet be drawn. As the author of one of the studies, I would like to conclude that the Hindley, Larner, and Kamerschen position is correct, but unfortunately, the empirical evidence at this time does not warrant such a conclusion. However, since my sample was larger and more comprehensive than that of Monsen and employed basically the same method of classifying firms as being owner or manager controlled, I think that Monsen and hence Mermelstein also must conclude that the issue is in doubt. While this in no way detracts from the overall value of Mermelstein's study, I would submit that his conclusions regarding the factors accounting for the largest corporations' *growing* advantages may have to be tempered in light of recent empirical evidence. I think this particularly applies in Mermelstein's case, for Larner's results indicated that the degree of control has no influence on either the *level* or the *variability* of profits, whereas Monsen found that the degree of control affected only the *level* of profits. And it is the *dispersion* rather than *level* that seems crucial in the Mermelstein study.

REFERENCES

R. J. Arnould, "Conglomerate Growth and Profitability," in L. Garoian, ed., *Economics of Conglomerate Growth*, Corvallis 1969.

A. A. Berle and G. C. Means, *The Modern Corporation and Private Property*, New York 1932; 2d ed. New York 1968.

N. R. Collins and L. E. Preston, "The Size Structure of the Largest Industrial Firms,

[4] Hindley (1970) states that he has obtained results "similar to those reported by Kamerschen" using "different methods and data." At the time the present communication was written, I did not have access to this article.

[5] They go on to say, "As is often the case, this led him to seek an explanation. He found that, although the members of the top management group need not be stockholders, they usually do hold sizable amounts of the stock in the corporation they manage, this stock often being acquired as a direct result of bonuses or compensation for their services. Most top managers of sucessful corporations are wealthy men, much of whose wealth is represented by stock in their own companies. For example, when Semon Knudsen left General Motors in 1968 to become President of Ford, he owned $3.3 million of G.M. stock. In addition, Ford gave him 15,000 shares of its common stock worth $850,000 and an option to buy 75,000 more shares at a price below the price the public has to pay for it. Larner's study showed the total income of managers to be closely related to the profitability of the companies they managed" (pp. 370–71). Since the time the above passage was written, Knudsen has also left Ford.

1909–1958," *Amer. Econ. Rev.*, Dec. 1961, *51*, 986–1011.

M. Hall and L. Weiss, "Firm Size and Profitability," *Rev. Econ. Statist.*, Aug. 1967, *49*, 319–31.

B. Hindley, "Capitalism and the Corporation," *Economica*, Nov. 1969, *36*, 426–38.

———, "Separation of Ownership and Control in the Modern Corporation," *J. Law Econ*, Apr. 1970, *13*, 185–221.

D. R. Kamerschen, "The Influence of Ownership and Control on Profit Rates," *Amer. Econ. Rev.*, June 1968, *58*, 432–477; correction Dec. 1968, 1376.

———, (1969a) "The Effect of Separation of Ownership and Control on the Performance of the Large Firm in the U. S. Economy," *Riv. Intern. di Scien. Econ. e Comm.*, July 1969, *61*, 293–301.

———, (1969b) "Recurrent Objections to the Theory of Imperfect Competition," *Zeitschrift für die Gesamte Staatswissenschaft*, Oct. 1969, *125*, 688–94.

———, "Conglomerate Mergers: The Myth and the Reality," *St. Johns Law Rev.*, Apr. 1970, *44*, 133–51.

A. D. H. Kaplan, *Big Enterprise in a Competitive System*, Washington 1954; rev. ed. Washington 1964.

R. J. Larner, "Ownership and Control in the 200 Largest Nonfinancial Corporations, 1929 and 1963," *Amer. Econ. Rev.*, Sept. 1966, *56*, 777–87.

———, "Separation of Ownership and Control and its Implications for the Behavior of the Firm," unpublished doctoral dissertation, Univ. Wis. 1968.

R. G. Lipsey and P. O. Steiner, *Economics*, New York 1966; 2d ed. New York 1969.

D. Mermelstein, "Large Industrial Corporations and Asset Shares," *Amer. Econ. Rev.*, Sept. 1969, *59*, 531–41.

R. J. Monsen and A. Downs, "A Theory of Large Managerial Firms," *J. Polit. Econ.*, June 1965, *73*, 221–36.

R. J. Monsen, J. S. Chiu, and D. E. Cooley, "The Effect of Separation of Ownership and Control on the Performance of the Large Firms," *Quart. J. Econ.*, Aug. 1968, *72*, 435–451.

G. J. Stigler, "The Statistics of Monopoly and Merger," *J. Polit. Econ.*, Feb. 1956, *64*, 33–40; reprinted in D. R. Kamerschen, ed., *Readings in Microeconomics*, Cleveland 1967; New York, 1969, 332–343.

———, *The Organization of Industry*. Homewood 1968.

Large Industrial Corporations and Asset Shares: Comment

By Stanley E. Boyle*

In a recent issue of this *Review*, David Mermelstein takes what, unfortunately, is not a very fresh look at the "share stabilization process on a decade-by-decade basis to find out whether the largest of these firms are better able to maintain their relative position during recent decades than those nearer the turn of the century" (p. 532).

While agreeing with the basic results obtained by Norman Collins and Lee Preston in 1961 and subsequently, Mermelstein argues that their work is marred by two serious shortcomings. First, the time periods employed by them (1909–19, 1919–29, 1929–35, 1935–48, and 1948–58) are of unequal length, making "meaningful comparisons difficult" (p. 532). Second, he believes that the approach they use "does not enable us to determine whether the type of stability increases or decreases the share of the very largest of the corporations being studied" (p. 532).

To offset the apparently overwhelming analytical problems raised by the Collins and Preston study, Mermelstein proposes to measure the changes in mobility exhibited by the 100 largest industrial corporations over the periods 1909–19, 1919–29, 1929–39, 1939–48, 1948–58, and 1958–64. Since some of the periods selected by Mermelstein are also of unequal length, the precise nature of this gain is unclear. Moreover, even if the periods selected were of equal length, it is questionable what may be gained from what is basically a repetition of the Collins and Preston "decade-by-decade" analysis. Analyses which involve such long periods ignore short-run changes in mobility and provide *no* insight into the timing and direction of changes which occur during a decade. As a consequence, any analysis of the effect of merger activity or antitrust policy which is based upon such flimsy evidence is of debatable value. For these and other reasons, those portions of the paper are excluded from the comment.

Four specific questions are analyzed in this note: First, what is the analytical significance of mobility changes which abstract from changes in the composition of the group? Second, what is the basic relationship between long-run (decade) and short-run (two year) mobility changes which occurred over the period 1919–64? Third, what questions are raised regarding the economic significance of analyzing mobility changes for any larger specified number of firms (100 or 200)? Fourth, what is the nature of the problems inherent to any suppositions regarding the extent to which overall mobility and changes in the level of aggregate concentration may be related?[1]

I. The Dropout Bias

"These [Mermelstein's] initial findings clearly suggest that increasingly the largest corporations are better able to hold onto their shares than their counterparts of an earlier era" (p. 534).

Mermelstein supports this basic conclusion with an analysis of changes in the rankings of *surviving* firms over a series of 10-year periods. The r^2 values obtained by Mermelstein compare the change in rank positions of these firms and are shown below as r^2_m (all survivors only) and r^2_{ma} (adjusted for industry shift). The values r^2_{bm} which show the com-

* Professor of economics, Virginia Polytechnic Institute. I owe an immeasurable debt to my former colleague, Professor Joseph P. McKenna, who is responsible for many of the computations which are used throughout this paper.

[1] The data employed in this note includes *only* manufacturing corporations and excludes the mining and trade companies included in the earlier studies. The collection of the basic data employed in this and an earlier paper (Boyle and Sorensen) was begun while the author served as Chief, Division of Industry Analysis, Federal Trade Commission. These differences should not affect the observations included in this comment since they are restricted to the method employed by Mermelstein rather than the actual statistical results obtained in his study.

TABLE 1—CHANGES IN ASSET-SIZE RANKINGS FOR ALL SURVIVORS (ADJUSTED AND UNADJUSTED) AND ALL FIRMS: 1909–58

Years Compared	r_m^2	r_{ma}^2	r_{bm}^2
1909, 1919	0.92285	.81045	—
1919, 1929	0.73704	.88654	0.407
1929, 1939	0.91929	.94472	0.678
1939, 1948	0.92686	.95555	0.691[a]
1948, 1958	0.93864	.93968	0.649[b]

[a] For the period 1939–49.
[b] For the period 1949–59.
Source: The r_m^2 and r_{ma}^2 value are taken from Mermelstein's *Review* article. The r_{bm}^2 are taken from Boyle and McKenna's forthcoming article, Table 1.

parable results obtained by Stanley Boyle and Joseph McKenna are computed on an "all firms" rather than an all survivors basis.[2]

It is interesting to note that the r_{ma}^2 values in Table 1 follow roughly the same path as do the r_{bm}^2 values. The most prominent difference between them is that they show substantially different levels. In both cases, however, the r^2 values are lowest for the 1919, 1929 comparison and then rise over the next two periods declining slightly in the last. The difference in level is, however, an exceedingly important property of the comparison. *If* the fact that there has been relatively little movement in rank of leading firms is economically significant, and presumably related to the structure of the industry, then the size of the segment of industry under examination is of considerable importance. Put another way, the decade of the 1920's was one of considerable mobility because 1) the rank of surviving firms changed, and 2) new firms entered and old firms departed from the charmed circle (the 100 largest).[3] This latter movement indicates that the rise of new competitors may well be more important than internal rank changes.

The r_m^2 and r_{ma}^2 values for the period 1919–29 are .737 and .887, respectively, compared with the r_{bm}^2 value of .407. The bulk of this difference is due to the fact that the latter value takes explicit account of the fact that by 1929 only 72 firms remained of the 100 largest of 1919. Carried to its most absurd extreme the "survivors only" approach for analyzing industry stability would produce an r_m^2 value of 1.00 if the leading firm alone among the top 100 retained its position of the period in question *but* the identity of all of firms ranked 2–100 changed over the period. Obviously, a value for $r_m^2 = 1.00$ in that situation is without economic significance. Therefore, the failure to note this second aspect of mobility, i.e., in and out of the group, and taking into account only the changes of those included in *both* periods constitutes a basic and substantial analytical shortcoming of the Mermelstein analysis.

II. The Decade Problem

As important as the preceeding problem is, its significance for competition is substantially outweighed by the fact that the measurement of mobility over a period as long as a decade ignores almost entirely the timing and direction of changes in large-firm stability. The differences between the paired-year and decade stability coefficients for the period 1919–39 are shown in Table 2.[4]

As would be expected, the data show greater stability (higher r^2 values) over the relatively short two-year periods than they do over the decade. More importantly, they show that mobility was the greatest between 1919 and 1923. Between 1923 and 1929 it remained almost constant, increased briefly between 1929 and 1931 and then declined sharply after 1931. Thus, examining only the decade movement obscures both the magnitude and timing of the significant movements which actually occurred between 1919 and 1931.

Looking at the paired-year values shown in Table 2 for the postwar period, it is possible to follow the progress of the large merger movement. Ignoring the period prior

[2] In addition to the differences in coverage, i.e., all survivors compared with all firms, the Boyle-McKenna data begin in 1919 rather than 1909, and actually pertain to decades, i.e., 1939–49 and 1949–59 rather than 1939–48 and 1948–58.

[3] Mermelstein does make an interesting contribution here in attempting to measure the importance of industrial shift upon rank stability. This was attempted in a much less precise matter by Boyle and McKenna for the period 1919–64 (pp. 10–14).

[4] The data for the entire period (1919–41 and 1948–64) are included in Table 2.

TABLE 2—PAIRED-YEAR AND DECADE STABILITY COEFFICIENTS: 1919, 1921 to 1962, 1964

Years Compared	Paired-Year Stability Coefficients	Decade Stability Coefficients	Time Period	Paired-Year Stability Coefficients
1919, 1921	.763		1939, 1941	.840
1921, 1923	.738			
1923, 1925	.800	.407	1948, 1950	.916
1925, 1927	.805		1950, 1952	.879
1927, 1929	.802		1952, 1954	.898
1929, 1931	.770		1954, 1956	.831
1931, 1933	.871		1956, 1958	.902
1933, 1935	.931	.678	1958, 1960	.859
1935, 1937	.886		1960, 1962	.845
1937, 1939	.907		1962, 1964	.841

Source: See Boyle and McKenna; Table 1.

to 1948 which was affected strongly by the economic changes growing out of World War II, we see that the highest degree of stability was shown between 1948 and 1950. Since then, however, with minor exception *apparent* mobility has increased.[5] The decade data alone fail to describe this pattern of steady change. Thus, the faith evidenced in, and the analytical conclusion drawn from decade averages such as those presented by Mermelstein must be tempered in light of their obvious shortcomings. Long-term stability data are useful *only* if proper attention is devoted the concommitant short-term changes which occur within the longer period.

III. The Relevant Size Class

A third serious, but scarcely unique problem of the Mermelstein paper surrounds its preoccupation with the changes which occur among the 100 largest industrial concerns. While the use of this level of firm aggregation may be of some value as an interim analytical device, one should remember that data for that group have no unique economic significance. There is little, if any, evidence that data for *any* particular number of firms has *any* critical economic value, as George Stigler pointed out in his critique of the Kaplan mobility findings more than a decade ago.

It is obvious that few, if any, economists, Stigler among them (1966, pp. 232–33), have paid sufficient attention to this admonition in recent years. The advice is, nonetheless, sound. Never put all of your eggs in one basket, particularly an irrelevant one: Two of the studies referred to above (see Boyle and McKenna, Boyle and Robert Sorenson) do analyze other levels. The second examines intra-industry mobility at the 10, 20, and 50 largest firm levels. Both of these papers show that two significant changes occur as the number of firms under examination expands. First, the level of the r^2 values tend to increase. Second, they tend to become somewhat more stable, i.e., they exhibit essentially smaller shifts between years.

In an effort to portray these changes over the period 1919 and 1964, the 100 largest firms have been subdivided into four groups (1–25, 26–50, 51–75, and 76–100) and paired-year stability coefficients computed for them. Table 3, shows the mean $\bar{x}$ and standard deviation σ of the paired year stability coefficients for each asset-size sub-group over each of the four major subperiods (1919, 1929; 1929, 1939; 1948, 1956; and 1956, 1964).[6]

The data presented in Tables 2 and 3 show that while the level of the stability coefficients for the 100 largest industrial concerns declined somewhat since 1948, that was not true for the 25 largest. On the contrary, the mean ($\bar{x}$) of stability coefficients for the subgroup of largest firms actually increased in each of the subperiods. At the same time the σ values declined. Clearly mobility declined for firms in the top size class. Movements in the other size classes show lower and more variable $\bar{x}$ values and larger values for σ. In some years the $\bar{x}$ values for the 76–100 size class group are zero are even negative. They show no significant

[5] Recent discussion of this phenomenon (Boyle and Sorensen) shows that "real" mobility (that which is not a function of changes wrought by merger activity) actually declined between 1950 and 1964 in the seven 2-digit industries which were examined.

[6] The paired-year coefficients for each of the component size classes are available from the author by writing department of economics, V.P.I., Blacksburg, Virginia 24061.

TABLE 3—COMPUTED MEAN $\bar{x}$ AND STANDARD DEVIATION σ OF PAIRED-YEAR STABILITY COEFFICIENTS BY SIZE CLASS

Size Class	1919, 1929[a]		1929, 1939[a]		1948, 1956[b]		1956, 1964[b]	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
1–25	.6544	.2445	.8989	.2068	.9520	.0161	.9622	.0067
26–50	.5567	.3053	.4820	.3023	.7550	.0502	.4569	.2344
51–75	.4324	.0824	.6886	.1847	.6106	.2289	.6007	.1414
76–100	.4497	.2126	.3164	.2874	.4522	.1263	.5607	.1707

[a] The values shown are based upon paired-year coefficients covering 5 time periods.
[b] The values shown are based upon paired-year coefficients covering 4 time periods.

pattern. On the average they seem to show smaller values for $\bar{x}$ but higher σ values.

The basic reason for this divergence in the $\bar{x}$ and σ value is the extent to which merger activity has affected the membership of each group. The approach used by Mermelstein to determine the impact of mergers (pp. 536–38) fails to take adequate account of the impact of the merger movement of the past decade and one-half.[7] As Boyle and McKenna pointed out: "An outstanding example of stability appears in the very largest firms. The 12 largest manufacturing firms in 1929 remained among the 12 largest firms in 1962, one-third of a century later" (p. 5).

While Mermelstein and others may have focused their analysis only upon changes among the 100 or 200 largest industrial corporations, and while these may be convenient numbers to use for the sake of presenting consistent estimates of the control over economic resources of some fixed number of firms through time, it should *not* be inferred that such estimates necessarily contain specific economic significance.

IV. Stability and Concentration

Although Mermelstein does not specifically direct his attention to the relationship between the level of aggregate concentration and the degree of stability which exists, it seems worthwhile to raise this question briefly at this time. Unfortunately, it is not one to which precise answers can be given. However, the data used in this comment taken in conjunction with those which appear in a recent Federal Trade Commission Report show an interesting relationship. The FTC study shows that the share of total corporate manufacturing assets held by the 100 largest firms increased from 39.4 percent in 1951 to 47.1 percent in 1958, holding at about that level until 1966. In 1968 and 1969 it rose again (p. 176).

Table 4 may serve to indicate the relationship between changes in stability and changes in the level of aggregate concentration for subgroups within the 100 largest over the period from 1950 to 1962. These data show that the increase in the level of concentration among the 100 largest corporations (5.5 percentage points) was the summary result of growth by the 10 largest firms (FTC p. 121). These data, taken in conjunction with those shown earlier (Boyle and McKenna), indicate strongly that when appropriate levels of aggregation are selected there may indeed be a strong correlation

TABLE 4—CHANGES IN CONCENTRATION, BY ASSET-SIZE CLASS WITHIN THE 100 LARGEST CORPORATIONS: 1950–1962

Size Class	Change in Concentration	Cummulative Change
	(percentage points of change)	
1–10	3.1	3.1
11–20	0.2	3.2
21–50	0.9	4.1
51–100	1.4	5.5

Source: Computed from, Willard Mueller, "Statement," p. 121. The figures used represent slight differences from those contained in the original table and are based upon a revision of those data.

[7] The reader interested in this area is directed to: FTC, Economic Report on Corporate Merger, particularly chapters 4 and 5.

between stability, i.e., the ability of old firms to maintain their relative position, and changes in concentration.

V. Conclusions

Although my comments have been numerous and rather pointed, it should not be inferred that Mermelstein's paper is without merit. Rather they should be interpreted as an attempt to indicate the type of problems which may ensue when economists and others are too long bound to the use of limited data sets which tend to obscure rather than explain significant changes through time.

REFERENCES

S. E. Boyle and J. P. McKenna, "Size Mobility of the 100 and 200 Largest U.S. Manufacturing Corporations. 1919–1964," *Antitrust Bull.*, forthcoming.

S. E. Boyle and R. L. Sorensen, "Concentration and Mobility: An Alternative Measure of Industry Structure," *J. Ind. Econ.*, forthcoming.

N. R. Collins and L. E. Preston, "The Size Structure of the Largest Industrial Firms, 1909–1958," *Amer. Econ. Rev.*, Dec. 1961, *51*, 986–1011.

A. D. H. Kaplan, *Big Enterprize in a Competition System*, rev. ed., Washington 1964.

D. Mermelstein, "Large Industrial Corporations and Asset Shares," *Amer. Econ. Rev.*, Sept. 1969, *59*, 531–41.

W. F. Mueller, "Statement," *Economic Concentration: Part I*, Hearings before the Senate Subcommittee on Antitrust and Monopoly, Washington 1964, 109–29.

L. E. Preston, "Statement," *Economic Concentration: Part I*, Hearings before the Senate Subcommittee on Antitrust and Monopoly, Washington 1964, 56–70.

G. J. Stigler, "The Statistics of Monopoly and Merger," *J. Polit. Econ.*, Feb. 1956 *64*, 33–40.

———, "The Economic Effect of the Antitrust Laws," *J. Law Econ.*, Oct. 1966, *9*, 225–37.

U.S. Federal Trade Commission, *Economic Report on Corporate Merger*, October 1969.

Large Industrial Corporations and Asset Shares: Reply

By David Mermelstein*

By bringing to my attention a body of data I evidently overlooked, David Kamerschen properly disputes my suggestion that differential patterns in managerial behavior —patterns relating to corporate control— may partially explain the increased ability during recent decades of the largest corporations to maintain their share of total assets.[1] Since Kamerschen does not challenge what I consider to be the major contribution of my article—its methodology and empirical findings—we are left, even more than before, with a shortage of possible valid explanations for the upward trend in the regression coefficients.

One possible explanation of my empirical findings, not mentioned in my previous article, may be found in the activities of the state. For example, substantial indirect subsidies have been given to the oil and automobile industries through construction, maintenance, and administration of a vast network of roads and highways.[2] Subsidies do not exist solely on the expenditure side of the state budgetary ledger. By permitting individuals to deduct mortgage interest payments from their taxable income, as well as state and local property taxes, and by not requiring individuals to report the imputed rental value of owner occupied homes, the state not only contributes to urban and suburban sprawl but decentralizes the population in such a way as to increase the demand for automobiles and hence indirectly subsidizes this industry.[3]

This argument has obvious merit in explaining my empirical results. It is not, however, without its limitations. Should all firms in an industry share in government subsidies equally (that is, in proportion to their size), then no increase should take place in those regression coefficients which have been adjusted for changes in industrial structure. To the extent that the largest firms receive a greater than average share of the complex mix of direct and indirect subsidies—a highly plausible hypothesis, to say the least—then some increase would occur in the regression coefficients over time, even when adjustments are made for changes in industrial structure.[4]

Turning next to the comment of Stanley Boyle, I am gratified that his research tends, by and large, to confirm my own. As Boyle himself points out, the values of r^2_{ma} and r^2_{bm} (in his Table 1) "follow roughly the same path." Moreover, his disaggregation of the data, presented in Tables 2–4, seems to be consistent with the work I have done and more in the nature of a complement than a substitute.

On the other hand, in the process of consistently presenting my research as though it were an analysis of changes in corporate *ranks* as well as in focussing on my parenthetically mentioned measuring rod of stability, r^2 the coefficient of determination, Boyle has unfortunately misrepresented my article and its somewhat different approach to industrial stability. Ranks, per se, were

* Polytechnic Institute of Brooklyn.

[1] Kamerschen is less than fair, however, when he claims that I failed to tie in my explanations "more explicitly" with the "enormous rise in conglomeration in recent years." I twice mention the effects of a recent increase in diversification and in the second of these discussions considerable space is allotted to this problem. In fn. 19 of my original article, I make it clear that in referring to diversification, I had conglomerateness largely in mind.

[2] A disproportionate number of the twenty largest firms have been in the oil and automobile industries since 1909:

1909	1	1929	10	1948	10	1964	10
1919	7	1939	10	1958	11		

Source: Mermelstein (1967, pp. 90, 97–150).

[3] For a brilliant and penetrating analysis of state expenditures in a capitalist society, see J. O'Connor (1970).

[4] The interested reader is referred to my dissertation for a discussion of other factors, such as the market for managers, bureaucratic practices, and dividend disbursements that may help to explain the upward trend of the regression coefficients.

not even studied and in analysis of what I did study—the relative shares of corporate assets—use was made of the regression coefficient, which in contrast to r^2, as explained in the original text, measures a different kind of stability from the latter. It is perfectly possible, for example, for r^2 to be higher during recent decades (or during Boyle's paired-years) with b yet remaining constant. This would indicate no increase, on average, in the ability of the largest firms to maintain their shares, but rather some decrease (as indicated by the higher r^2) in erratic changes in shares (or ranks).

Because he does not notice the differences between an analysis based on comparisons of regression coefficients and one based on coefficients of determination, Boyle errs when he asserts that my "survivors only" method has a "basic and substantial shortcoming." My attempts to check out the biases involved in this procedure show it to be, most probably, a reliable and unbiased method, *given the uses to which it was put*. Boyle is of course correct, though hardly relevant, in pointing out that a single firm would have an r^2 of unity. In actuality, the number of survivors for the later periods stabilized in the mid-1930's, and even the number in the earliest and most volatile decade remained relatively high at 61: 1909-1919, 61; 1919-1929, 70; 1929-1939, 85; 1939-1948, 84; 1948-1958, 85; and 1958-1964, 88. A more extended argument that the "survivors only" method is of negligible bias, I relegate to the Appendix.

Elsewhere in Boyle's comment, and in Kamerschen's as well, a question is raised concerning the relevance of certain kinds of research engaged in by students of industrial organization. Both (independently) approve George Stigler's statement that "The statistical universe of the hundred largest corporations is inappropriate to studies of monopoly and competition, and we may hope that this (referring to Kaplan) will be the last study to fall prey to its dramatic irrelevance."[5] Research on "the extent of instability in the relative fortunes of the leading firms," Stigler continues, should be "applied on a correct, industry basis." (In passing, if Stigler is correct, then Kamerschen's statement that "... Mermelstein has made an important contribution to the empirical literature of industrial organization" is gratuitous.)

Stigler's viewpoint, as expressed in the passage quoted by Boyle and Kamerschen, is of considerable importance.[6] Behind it lie key assumptions about the nature of the capitalist process and the relationship between economic phenomena and the political institutions within which economic decision making takes place in this country. Many economists share Stigler's sentiments; nonetheless, I believe he is mistaken. Without, I hope, violating the spatial privileges given an author to reply, I would like to set down a few introductory comments suggesting an alternative framework in which these issues can be judged.[7]

To begin, it is not self-evident that conglomerateness, and by the same token, the statistical universe of the one or two hundred largest industrial corporations, has no relation whatsoever to economic power in the marketplace. For example, Corwin Edwards in a statement to a Senate Subcommittee lists four ways in which conglomeration enhances market power: a) subsidization; b) reciprocity; c) full line selling; and d) the forbearance that prevails among large conglomerates (pp. 43–45). Mere size alone may

[5] In the interests of clarity, it should be pointed out that Kamerschen errs in implying that I, like A. D. H. Kaplan, use the "turnover" method and have not "satisfactorily answered the 'hostile comments and reviews'" [of Kaplan's book]. To the contrary, my own approach to this aspect of industrial stability uses what I consider a more refined statistical *alternative* to the turnover method used by Kaplan. Kamerschen indicts my study by declaring it to have "theoretical limitations." Let us be clear that these charges do not apply to the area of method at all, either Kaplan's turnover process or my regression analysis, but to the subject matter itself—the universe of the largest corporations

[6] Another leading spokesman of this point of view, Morris Adelman, in a statement before a Senate Subcommittee, has asserted that "absolute size is absolutely irrelevant" and by implication that neither conglomeration nor increased overall concentration can ever have any harmful competitive effects (p. 228).

[7] Many of these comments are developed at greater length in the introductory material and a number of the readings in Mermelstein (1970).

also be instrumental in obtaining less expensive credit[8] or in perpetuating industrial control of various markets through the ability to purchase the pertinent technology. More important, absolute size has always been thought to be crucial in times of war. This time honored precept applies not merely to nations but to industrial corporations as well. Nor need wars break out for power to be exercised: threat of war by itself is often sufficient.

It is interesting to ponder the effects on pricing, output, resource allocation, and technological development if there existed a total of only one hundred firms in the entire economy, each firm of equal size, and each operating in every existing industrial market. With (4-firm) concentration ratios of 4 percent, this economy, by Stigler's judgment, must be super-competitive. Remembering Edwards' discussion, it would be rash to assert that there would be absolutely no difference between this situation and one in which each industry contained one hundred equal-sized firms, not one of which operated in more than one industry.

To pursue the matter one step further, have we collected industrial data and fashioned our definitions of industry with such theoretical precision that it can be said with perfect assurance that studies of overall concentration have only "dramatic irrelevance?" The current merger movement[9] contains an extraordinary number of firms desperately striving to avoid dependence on a single industrial market. John Galbraith's *The New Industrial State*, whatever its failings,[10] correctly perceives that a substantively new pattern of industrial organization has been created, or at least is in the process of being developed, dissimilar only in degree from my hypothetical example above (where total output was produced by one hundred firms). What is happening sorely needs further study, not casual dismissal out of hand.

The primary reason for concern is that power in the market place is but one facet of industrial concentration. Power also exists to influence or control key governmental decisions; or, in other words, industrial concentration has a political dimension of vast importance.[11] It is no accident, for example, that Presidents have been prone to appoint bankers, businessmen, and Wall Street lawyers to cabinet posts and other policy-making positions inside the executive branch of government.[12] This power is continually exerted in such areas as foreign policy, military contracts, labor, taxation, budgetary programs, urban policies, transportation, and basic research.

As an example, take two problems very much in the news these days—air pollution and urban mass transit. No one can question their importance: the condition of the latter is part of the breakdown in our urban centers while the former is beginning to threaten the life process itself.[13] Yet nothing much is done. This should come as no surprise since lasting solutions to these and other problems of American society require radical changes of vast magnitude. At the very least, we need an all-out attack on the population problem, rejection of economic growth as a way of life,[14] and comprehensive regional planning involving the location of jobs, housing, and

[8] Donald Dewey's argument (quoted in my original article), that advantages in the capital market may be due to the government aid that giant firms are likely to receive when threatened by bankrupcy, is apropos in this context.

[9] For a good discussion, see Reid pp. 73–120.

[10] For example, Galbraith shows little interest in the overseas activities of the American corporations and in the rise of the multi-national corporation. See also R. Miliband, reprinted in Mermelstein (1970).

[11] For a classic discussion of some of the political aspects of oligopolistic rivalry, see K. W. Rothschild. As Rothschild points out, "The fact is that when we enter the field of rivalry between oligopolistic giants, the traditional separation of the political from the economic can no longer be maintained. . . . *The oligopolistic struggle for position and security includes political action of all sorts right up to imperialism. The inclusion of these 'non-economic' elements is essential for a full explanation of oligopoly behavior and price*" (pp. 462–63, as reprinted in *Readings in Price Theory*; italics in original).

[12] Evidence that American society is ruled by an upper class based on possession of wealth is contained in G. W. Domhoff.

[13] For a sober appraisal, see B. Commoner.

[14] What follows, by author Edward Abbey (quoted by Chapman and Harrington), is the best succinct statement I have come across expressing by analogy the crux of the problem: "Growth for the sake of growth is the ideology of the cancer cell."

recreation in which the real needs[15] of the people are serviced.[16]

Unfortunately, these changes are in sharp conflict with existing property rights and the way in which corporations have defined their economic interest as well as their preferred modes of operation. Consequently, pollution, congestion, deteriorating mass transit, as well as the other social ills we associate with urban blight, do not just happen. They are the end products of a chain of decisions the first links of which have been forged by those at the apex of the corporate system. Increasingly, economic growth—the paramount goal of American life[17]—has been spearheaded by the largest of these corporations. Their decisions have structured American society. Their vision is one of continued expansion of one kind of consumption—goods sold in the marketplace for profit—giving short shrift not only to the alternatives of public goods and greater leisure but to the social and ecological repercussions as well. This vision has molded the consciousness of the American people, and while the corporations give lip service to the ideology of the free market, they do not hestitate to use political means to perpetuate their ideological and economic hegemony.

The crisis in pollution and transportation is not unique. Whether the issue is crime or war, miseducation or addiction, poverty or racism, solutions are nowhere in sight. Instead, we have a vast network of government subsidies, government controls, and government programs of all kinds, many in conflict with each other, but virtually all in service of one or another industrial interest.

Corporate capital has never been so powerful as it is today,[18] given the relative displacement of other propertied classes such as the independent farmer and the small businessman.[19] Increasing concentration of assets, more secure than ever, serves to ease the task of coordinating decisions in defense of corporate hegemony, a task of some importance for those who occupy the upper reaches of the corporate world, given the current situation in America, one in which growing numbers of Americans, especially its youth, are deeply alienated and openly hostile to the capitalist way of life.

Appendix

In the September 1969 issue of this *Review*, I used the following regression procedure to determine mobility of large corporations. For each time period, the share of survivor assets is regressed on their corresponding early year shares. A regression coefficient greater than unity means

[15] Without minimizing the immense difficulties involved in defining this term, I would suggest a meaningful approach is to be found in the social and biological requirements needed to preserve the species: nutrition, protection from environmental hazards, reproduction, and emotional well-being. (I am indebted to my colleague, Professor Shane Mage, for this approach. Helpful insights about the meaning of human needs have also come from a variety of sources, especially Galbraith, P. Baran, P. Baran and P. Sweezy, and H. Marcuse.)

[16] Redirection of social priorities and productive energies may be the beginning of the process by which we replace the existing set of status symbols—those which emphasize the acquisition of material wealth—with an altogether different set of social standards in which special status, to the extent that it continues to exist at all, is based on social responsibility, service to the community, and warm, decent, and honorable relationships with one's fellow human beings.

[17] Nor is it likely that a capitalist economy could function otherwise. A zero rate of growth means either that property owners must consume their profits, a difficult task to say the least, given the existing distribution of income—and one not likely to receive social sanction—or that profits be distributed directly in the form of gifts or loans or indirectly in the form of lower prices in which case no rationale remains for continuation of a private ownership system.

[18] A. A. Berle (p. 102), a former State Department official and professor of corporation law at Columbia University, does not exaggerate when he writes:

> But in terms of power, without regard to asset positions, not only do 500 corporations control two-thirds of the non-farm economy but within each of that 500 a still smaller group has the ultimate decision-making power. This is, I think, the highest concentration of economic power in recorded history. Since the United States carries on not quite half of the manufacturing production of the entire world today, these 500 groupings—each with its own little dominating pyramid within it—represent a concentration of power over economics which makes the medieval feudal system look like a Sunday school party. In sheer economic power this has gone far beyond anything we have yet seen.

Berle, of course, draws other, more conservative conclusions than I from this perception of social reality.

[19] See J. O'Connor (1968) for a more elaborate statement of this argument and its implications.

that initially large firms achieved an even larger share in the later year.

To facilitate an analysis as to whether biases have crept into my procedures, I have classified those firms that exit from the ranks of the largest 100 by the ranks they previously had held and entrants by the ranks they achieve. For 1909–1919, we find that of those firms ranked 1–50 in 1909, 19 were ranked 51 or smaller in 1919. Of these, 5 were "exiters" while the other 14 managed to survive in the lower half of the 1919 listings. As many as 34 firms, ranked 51–100 in 1909, dropped out of the largest 100 by 1919. Thus, most of the non-surviving firms of 1909–1919 were of small or medium size. Taking into account only the "exiting" firms of 1909–1919, it would appear that the survivors only method understates the regression coefficient. This can be explained as follows: most of the 1909 firms—14 of 19—which fall out of the ranks of the largest 50 affect the regression since they remain among the survivors. None of the 36 smaller firms which drop into ranks 101–150 affect the regression. The regression coefficient is therefore *understated* since the relative declines of the large firms are counted while those of the smaller firms are not.

In contrast, smaller firms not ranked in 1909 but ranked among the largest 100 of 1919 often made gains relative to the larger firms. Fifteen of the 39 entrants of 1909–1919 achieved a position in the first 50 ranks. To the extent that these gains of the relatively small do not register in the 1909-1919 regression, the regression coefficient may be *overstated*. We therefore conclude that the survivors only method unavoidably introduces a set of biases for the period 1909–1919 of undetermined direction and magnitude, but they appear to work at cross purposes and thereby tend to cancel each other out.

The situation as to entries and exits, as revealed in Table 1, changes little between the first decade and the second, but after 1929, the picture is considerably different. Rarely does an entrant rise into the ranks of the largest 50; rarely does a member of the largest 50 fall out of the ranks of the largest 100. As a result, a small downward bias probably exists; that is, the late coefficients are somewhat lower than they should be. The reason is somewhat complicated. When exiters are small, the gains of the large firms vis-à-vis the small exiters are not understated. In contrast, the gains made by the entrants are at the expense of the smaller firms—those ranked 50 or higher—and not at all at the expense of the largest corporations. For this reason, the regression coefficient is relatively unaffected. To the extent the above analysis is correct, it is possible that we may have slightly understated the increased ability of the large firms to maintain their shares over the years 1909–1964.

What we are trying to determine is whether the survivors only method which uses most of the 100 largest firms, but not all of them, cause *b* the regression coefficient to be either understated or overstated. This problem tends to diminish as more and more firms are used as observations. Since this is true, we can look at the path taken by the regression coefficients as we move from a study

TABLE 1—EXITING AND ENTERING FIRMS, CLASSIFIED BY RANKS

		Ranks 1–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100	Total
1909–1919	Exiters	0	1	0	2	2	4	7	9	8	6	39
	Entrants	4	1	3	4	3	6	4	3	5	6	39
1919–1929	Exiters	1	1	1	2	2	3	5	3	6	6	30
	Entrants	0	2	1	4	3	3	6	3	5	3	30
1929–1939	Exiters	0	1	0	0	3	0	3	2	2	4	15
	Entrants	0	0	0	0	0	1	2	1	5	6	15
1939–1948	Exiters	0	0	0	0	0	0	1	6	4	5	16
	Entrants	0	1	1	0	1	1	1	4	3	4	16
1948–1958	Exiters	0	0	0	0	0	0	1	3	6	5	15
	Entrants	0	0	0	2	2	2	0	6	1	2	15
1958–1964	Exiters	0	0	0	0	0	0	0	2	4	6	12
	Entrants	0	0	0	1	0	0	1	1	3	6	21

TABLE 2—REGRESSION COEFFICIENTS

	1909–1919	1919–1929	1929–1939	1939–1948	1948–1958	1958–1964
All survivors	0.73293	0.65403	0.96479	0.92046	1.11028	0.97590
First 75	0.75447	0.60347	0.97295	0.91434	1.08627	0.96163
First 50	0.76234	0.57467	0.96977	0.89558	1.13776	0.97254
First 25	0.79053	0.54448	0.92041	0.85448	1.13080	0.91898

TABLE 3—REGRESSION COEFFICIENTS, ADJUSTED FOR CHANGE IN INDUSTRIAL STRUCTURE

	1909–1919	1919–1929	1929–1939	1939–1948	1948–1958	1958–1964
All survivors	0.77094	0.84343	0.95668	0.92068	1.06927	0.96762
First 75	0.75591	0.79136	0.98028	0.90512	1.05739	0.95544
First 50	0.76675	0.75349	0.99998	0.90908	1.11953	0.88738
First 25	0.90691	0.73752	0.96544	0.90825	1.11928	0.92386

based on survivors from the ranks of the leading 25 firms to one based on survivors from the leading 100. In Tables 2 and 3 the various regression coefficients are displayed in a form suitable for a comparison of these differences.

The data indicates that both the adjusted as well as the unadjusted regression coefficients for all survivors may be somewhat overstated in 1909–1919 and understated in 1919–1929; for the decades 1939–1948 and 1958–1964, there appears to be a slight understatement, while for 1948–1958 a slight overstatement. No bias one way or the other appears in the remaining decade. This information does not conflict with our earlier judgments; if anything, it confirms them. We had suggested that the bias was indeterminate for the first two decades and our "trend of the coefficients" estimate of the bias shows that in later decades the bias, if any, tended towards understatement of the regression coefficient. The above data shows that in two of the three decades for which a trend exists, the coefficient is slightly understated. With so few observations, too much should not be made of these results. All things considered, the bias introduced by my procedure is probably negligible and of limited significance. Still, to the extent we can check on the direction of the bias inherent in the survivors only method, the major conclusion of my article, that the regression coefficients have exhibited a distinct upward trend indicating increased ability of the large corporations to maintain their shares, goes unchallenged. In fact, it is reinforced, given the lowered coefficient for 1909–1919 and the raised coefficient for 1958–1964.

REFERENCES

P. Baran, *The Political Economy of Growth*, New York 1957.

——— and P. Sweezy, *Monopoly Capital: An Essay on the American Economic and Social Order*, New York 1966.

A. A. Berle, "Economic Power and the Free Society," in A. Hacker, ed., *The Corporation Take-Over*, New York 1964.

D. Chapman and P. Harrington, "Land Lovers," *Look*, November 4, 1969, *33*, 54–61.

B. Commoner, *Science and Survival*, New York 1963.

G. W. Domhoff, *Who Rules America?* Englewood Cliffs 1967.

J. K. Galbraith, *The Affluent Society*, Boston 1958.

———, *The New Industrial State*, Boston 1967.

A. D. H. Kaplan, *Big Enterprise in a Competitive System*, Washington 1954; rev. ed., Washington 1964.

H. Marcuse, *An Essay on Liberation*, Boston 1969.

———, *One-Dimensional Man*, Boston 1964.

D. Mermelstein, *Economics: Mainstream Readings and Radical Critiques*, New York 1970.

———, "Large Industrial Corporations and Asset Share Maintenance, 1909–1964," unpublished doctoral dissertation, Columbia Univ. 1967.

——, "Large Industrial Corporations and Asset Shares," *Amer. Econ. Rev.*, Sept. 1969, *59*, 531–41.

R. Miliband, "Professor Galbraith and American Capitalism," in R. Miliband and J. Saville, eds., *The Socialist Register 1968*, London 1968.

J. O'Connor, "The Fiscal Crisis of the State," Part I, *Socialist Revolution*, Jan.-Feb. 1970, *1*, 14–53; Part II, Mar.-Apr. 1970, *2*, 39–94.

——, "The Situation at Present and What Is To Be Done?" *Mid-Peninsula Observer*, June 3–17, 1968, 9–12; reprinted under the title, "Some Contradictions of Advanced U.S. Capitalism," in D. Mermelstein, ed., *Economics: Mainstream Readings and Radical Critiques*, New York 1970.

S. R. Reid, *Mergers, Managers, and the Economy*, New York 1968.

K. W. Rothschild, "Price Theory and Oligopoly," *Econ. J.* Sept. 1947, *57*, 299–320; reprinted in K. E. Boulding and G. J. Stigler, eds., *Readings in Price Theory*, Homewood 1952.

U.S. Congress, Subcommittee on Antitrust and Monopoly of the Committee on the Judiciary, U.S. Senate, *Economic Concentration*, Hearings, Pt. 1, 88th Cong., 2nd sess., Washington 1964.

Welfare Aspects of a Regulatory Constraint: Note

By EYTAN SHESHINSKI*

Government agencies commonly employ the "fair rate of return" criterion in the regulation of monopolies: after the firm subtracts its operating expenses from gross revenues, the remaining revenue should be just sufficient to compensate the firm for its investment in plant and equipment, at a rate which is considered to be fair.

It has been argued by Harvey Averch and Leland Johnson, and now rigorously proved by Akira Takayama, that such constraint induces the firm, subject to regulatory control, to increase its investment and output and also to deviate from the optimal allocation of inputs, because the regulated firm does not equate marginal rates of factor substitution to the ratio of factor costs. Therefore, cost is not minimized at the output selected by the firm.[1]

Since the fair rate of return criterion leads to a nonoptimal state in the sense of Pareto, a basic question is whether it improves the performance of the economy, from a welfare point of view, as compared with the unregulated monopoly situation (where output is too small). This is a "second best" problem in which we have to choose between two situations, each deviating in one way or another from optimality.

Here we show that from the point of view of efficiency, disregarding income distribution aspects, some regulation via the fair rate of return is always advantageous. We also derive the rule for the optimal degree of regulation, i.e. the regulation that maximizes social welfare.

I. The Takayama Model

Consider a monopoly employing two inputs, capital K and labor L to produce a homogeneous output Y. Output is a function of inputs

$$(1) \qquad Y = f(K, L)$$

Each factor has a positive and decreasing marginal product. The marginal rate of substitution between inputs is decreasing, i.e., isoquants are concave.

The price of the product P is negatively related to the level of output by the inverse demand function

$$(2) \qquad P = P(Y), \qquad P'(Y) < 0$$

The costs of the inputs, r for capital and w for labor, are fixed for the firm. Total costs, C, are therefore $C = rK + wL$. The profit of the firm Π, is defined as

$$(3) \quad \Pi = PY - C = PY - rK - wL$$

Since depreciation is assumed away, the operating expenses of the firm are only labor costs. The fair rate of return criterion imposes the following constraint on the firm: Denote by s the fair rate of return determined by the regularity agency. The firm's net revenues (gross revenues minus operating expenses) per unit of capital should not exceed s

$$\frac{P \cdot Y - wL}{K} \leq s$$

or

$$(4) \qquad P \cdot Y - sK - wL \leq 0$$

The firm attempts to maximize (3) subject to (4). Under suitable assumptions about the profit function, and provided the solution is interior, the first-order necessary conditions for a maximum are[2]

$$(5) \quad (P + P'f)f_1 - r - \lambda[(P + P'f)f_1 - s] = 0$$

* Hebrew University, Jerusalem.

[1] This deviation from the optimal allocation of inputs persists even under a more flexible scheme for a "graduated fair return" criterion, as suggested by Alvin Klevorick, although the degree of inefficiency is clearly reduced.

[2] A rigorous analysis is given by Takayama.

$$(6)\quad (P+P'f)f_2-w \\ -\lambda[(P+P'f)f_2-w]=0$$

λ being the Lagrange multiplier.

When s exceeds the maximum rate of return earned by the unregulated monopoly, $\bar{s}$, the constraint is not binding, λ is zero, and we have from (5) and (6) the standard profit maximizing condition of a monopolist. At the other extreme, when s is less than r, profits are negative and the firm will prefer to shut down. In the border case $s=r$ (and $\lambda=1$) the firm is indifferent as between shutting down and operating. Thus, the realm of interest is that in which $0<\lambda<1$, where the constraint is effective ($\lambda>0$) and the fair rate of return exceeds the market cost of capital ($\lambda<1$).

At the constrained maximum point, production is inefficient, since from (5) and (6)

$$(7)\quad \frac{f_1}{f_2}=\frac{r}{w}-\frac{\lambda}{1-\lambda}\left(\frac{s-r}{w}\right)<\frac{r}{w}$$

the marginal rate of substitution between inputs is lower than the ratio of input prices. Each output is produced with more capital and less labor as compared to the unregulated optimum. This inefficiency derives from the fact that the net return of the firm on every unit of capital is $s-r$, and this creates an incentive to increase capital intensity.[3]

In the region in which $0<\lambda<1$, we have from (6)

$$(8)\quad (P+P'f)f_2-w=0$$

which is the standard profit maximizing rule of a monopoly for labor input. Equation (8), together with the constraint (4)

$$(9)\quad P\cdot Y-sK-wL=0$$

determine the constrained inputs (K, L) for any given level of s. We now wish to analyze the response of the firm to different levels of regulation.

Differentiating the profit constraint (9) with respect to s,

$$(10)\quad (P+P'f)\left(f_1\frac{dK}{ds}+f_2\frac{dL}{ds}\right) \\ -K-s\frac{dK}{ds}-w\frac{dL}{ds}=0$$

From (8) this reduces to

$$(11)\quad [(P+P'f)f_1-s]\cdot\frac{dK}{ds}=K$$

Since $s>r$ and $0<\lambda<1$, we have from (5) that $(P+P'f)f_1-s<0$, so (11) yields that $dK/ds<0$. The response of L to changes in s is found by differentiation of (8). Denoting the marginal revenue function by $h=P+P'f$, we have

$$(12)\quad [h'f_1f_2+hf_{21}]\cdot\frac{dK}{ds} \\ +[h'f_2^2+hf_{22}]\cdot\frac{dL}{ds}=0$$

It is assumed that marginal revenue is decreasing, $h'<0$, so the sign of the second-term in brackets is negative. Since $dK/ds<0$, dL/ds is negative provided the first-term in brackets is positive:

$$(13)\quad h'f_1f_2+hf_{21}>0$$

Let us make this assumption, which is equivalent to assuming that capital and labor are complement inputs.[4]

Thus, as regulation tightens, i.e., as s decreases, both inputs and output increase.

$$-\frac{dK}{ds}>0,\qquad -\frac{dL}{ds}>0,\qquad -\frac{dY}{ds}>0$$

II. Welfare Implications

Suppose the economy consists of identical individuals, so there is no income distribution problem. The utility, or social welfare function, is

$$(14)\quad U=U(Y,K,L)$$

The marginal utility of Y is positive,

[3] In the constrained region $\Pi=P\cdot Y-rK-wL=P\cdot Y-sK-wL+(s-r)K=(s-r)K$

[4] It can be shown that $\partial K/\partial w<0$ and $\partial L/\partial r<0$. Inputs are normal and complements provided (in addition to the other assumptions), $\partial^2\pi/\partial K\partial L=h'f_1f_2+hf_{21}>0$.

$U_1>0$, while the marginal utilities of capital and labor are negative, $U_2<0$ and $U_3<0$, reflecting the loss of forgone consumption and leisure, respectively.

The change in utility as a result of a change in s is

$$(15)\quad \frac{dU}{ds} = U_1\left(\left(f_1\frac{dK}{ds} + f_2\frac{dL}{ds}\right) + U_2\frac{dK}{ds} + U_3\frac{dL}{ds}\right.$$

The first-order conditions of the consumer who maximizes utility are

$$(16)\quad \frac{U_1}{P} = -\frac{U_2}{r} = -\frac{U_3}{w}$$

Substituting in (15) we get

$$(17)\quad \frac{dU}{ds} = \frac{U_1}{P}\cdot\left[(Pf_1 - r)\frac{dK}{ds} + (Pf_2 - w)\frac{dL}{ds}\right]$$

Let us determine the sign of dU/ds at the unconstrained point: $s=\bar{s}(\lambda=0)$. At this point, from (5) and (6)

$$Pf_1 - r = -P'ff_1$$

and

$$Pf_2 - w = -P'ff_2$$

Substituting in (17) we have

$$(18)\quad \frac{dU}{ds} = -\frac{U_1}{P}P'f\cdot\left(f_1\frac{dK}{ds} + f_2\frac{dL}{ds}\right) < 0$$

Decreasing s from the ineffective level $\bar{s}$, always raises utility. Therefore, *some regulation via the fair rate of return criterion is always worthwhile.*

III. Optimal Degree of Regulation

Since regulation can always improve welfare, it is interesting to find the level of s that maximizes utility. In the constraint region $(r<s<\bar{s})$, the necessary condition for a maximum of U is to set (18) equal to zero. This condition can be rewritten

$$(19)\quad \frac{dU}{ds} = \frac{U_1}{P}\cdot\left[P\frac{dY}{ds} - \frac{dC}{ds}\right] = 0$$

or

$$(20)\quad P\frac{dY}{ds} = \frac{dC}{ds}$$

where C is total costs. Condition (20) has an obvious interpretation: as s decreases, output and costs increase. As we have seen, the increase in costs stems both from the increase in output and from the inefficient combination of inputs.

It is desired to reduce s so long as the resulting value of the increased output exceeds the corresponding increase in costs, and vice versa. At the optimum point, the change in the value of output exactly equals the change in costs. Put in a different way, (20) can be rewritten as

$$(21)\quad P = \frac{dC}{dY}$$

which resembles the standard optimality rule of price equal to marginal costs, only that dC/dY is not the curve derived from minimum cost allocation of inputs.

In Figure 1, MC is the marginal cost curve

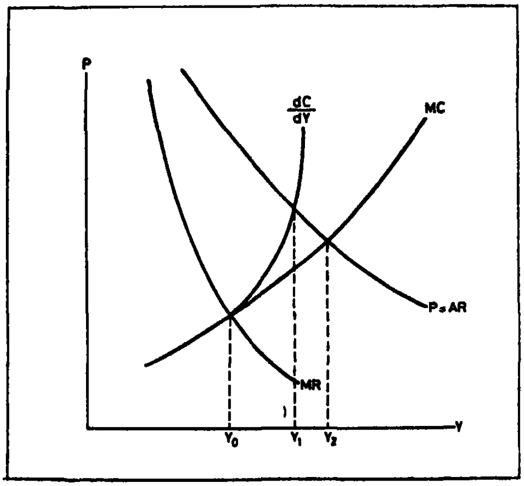


FIGURE 1

of the *unconstrained* monopoly, P the demand curve and MR the marginal revenue curve. The unconstrained optimal output is Y_0. Now, as s decreases from s, output and costs increase. The change in costs for a unit change in output is drawn by the curve dC/dY which is higher throughout than the MC curve (which represents minimum combination of inputs). The optimal s is set at the level corresponding to the point where dC/dY intersects the demand curve, at output Y_1.[5] The optimal output is always higher than that produced by the unconstrained monopoly (Y_0), and lower than the Pareto optimum output (Y_2).

[5] A comment on the "graduated fair return" criterion, suggested by Klevorick, might be in place here. He proposes that the fair return be a function of the amount of capital that the firm employs, rather than a fixed number. If one can choose the function optimally, a graduated maximal return can clearly not be worse than a fixed one, since the latter is a special case of the former. While this much is obvious, the index chosen by Klevorick, E, which indicates the deviation of the marginal rate of substitution between factors and the ratio of their prices, has no direct welfare implications. In fact, the absolute deviation of E from unity (Pareto optimality) is monotonically increasing as s decreases, but we already know that not all reductions in s necessarily improve welfare.

REFERENCES

H. Averch and L. L. Johnson, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1052–69.

A. K. Klevorick, "The Graduated Fair Return: A Regulatory Proposal," *Amer. Econ. Rev.*, June 1966, *56*, 477–84.

A. Takayama, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.*, June 1969, *59*, 255–60.

Pitfalls in Financial Model Building: Some Extensions

By Mark L. Ladenson*

In a recent contribution to this *Review*, William Brainard and James Tobin presented a general equilibrium model of the financial sector of the economy and a general disequilibrium model of the dynamic process by which the endogenous variables in the model adjust from one set of equilibrium values to another in response to a change in one or more exogenous variables. The variables in the models are linked by behavioral equations and by identities. The identities imply a number of restrictions on the coefficients of the variables in the behavioral equations and in general, imply that the coefficients of a particular variable summed across the behavioral equations add up to zero. This was one of the authors' basic points but the argument was presented without any formal derivations of the restrictions. Furthermore, the authors did not estimate the coefficients of the equations of their model but assumed certain values for the parameters and performed simulation experiments. In Section I of the present essay we give a formal derivation of the key restrictions of Brainard and Tobin's model, and in Section II we discuss some problems associated with estimating the parameters of the model.

I

Brainard and Tobin's static equilibrium model includes a sector containing financial assets and debts of the public and a sector involving bank asset holdings. In deriving the restrictions emphasized by the authors, however, it is unnecessary to deal with the banking sector. Brainard and Tobin assume that the public holds its net wealth in the form of five different types of financial assets and liabilities: demand deposits, time deposits, treasury securities, loans from banks, and equities. The demand for each of these is a linear function, homogeneous in net worth, of interest rates and national income:

$$(1) \qquad y_t^* = (B_1X_1 + BX_t)W_t$$

where X_1 is a variable which always takes on a value of unity, X_t is a 6x1 vector of interest rates and national income, B_1 is the 5x1 vector of coefficients of X_1 (constant terms), B is the 5x6 matrix of structural coefficients, W_t is net worth, a scalar, and y_t^* is the 5x1 vector of the values of the financial assets and liabilities given by current values of X_t and W_t.

The static equilibrium model of public behavior contains the system of five equations (1), and implicitly it also contains an *ex post* and an *ex ante* identity. Define the vector:

$$r' \equiv [1, 1, 1, -1, 1]$$

The *ex post* identity is

$$(2) \qquad r'y_t = W_t$$

where y_t is a 5x1 vector of actual values of the financial assets and liabilities of the public sector.[1] Equation (2) is simply a balance sheet identity; the sum of assets minus liabilities equals net worth.

The *ex ante* identity

$$(3) \qquad r'y_t^* = W_t$$

is a sort of "rational desires" hypothesis. It constrains the *desired* or equilibrium values of the financial assets and liabilities to obey the balance sheet identity and is analogous to the budget constraint in the familiar static theory of consumer behavior.

* Assistant professor of economics, Michigan State University. The chapter of my doctoral dissertation on which this paper is based was written with major assistance from Professor Walter D. Fisher. Others whose help and influence are reflected in the paper are Professors Frank Brechling, Patric Hendershott, Jan Kmenta, James Ramsey, and Arthur Treadway. While my deep thanks go to them, I remain solely responsible for errors.

[1] The fourth element in the vector y_t, loans from banks, is the only liability of the public; hence the negative value of the fourth element of r'.

$$
(4)\quad \begin{array}{l} y_{1t} - y_{1,t-1} = \alpha_{11}(y^*_{1t} - y_{1,t-1}) + \ldots + \alpha_{15}(y^*_{5t} - y_{5,t-1}) + \gamma_{1\Delta W}\Delta W_t \\ \vdots \\ y_{5t} - y_{5,t-1} = \alpha_{51}(y^*_{5t} - y_{5,t-1}) + \ldots + \alpha_{55}(y^*_{5t} - y_{5,t-1}) + \gamma_{5\Delta W}\Delta W_t \end{array}
$$

Brainard and Tobin go on to specify how this system behaves out of equilibrium with the set of adjustment equations (4), where the α_{ii} are own-adjustment coefficients, the α_{ij} $(i \neq j)$ are cross-adjustment coefficients, the $\gamma_{i\Delta W}$ are structural coefficients, and the subscript on y indexes a particular financial asset (or liability). The system (4) is a general disequilibrium system, a generalization of the familiar stock-adjustment model. The change in each financial asset depends not only on the gap between its own desired and actual values but on all such gaps and also on the change in wealth in the period. Equation (4) may be expressed more compactly as

$$
(5)\quad y_t - y_{t-1} = A(y^*_t - y_{t-1}) + \gamma_{\Delta W}\Delta W_t
$$

where A is the 5x5 matrix of adjustment coefficients, α_{ij}, and $\gamma_{\Delta W}$ is the 5x1 vector of structural coefficients, the ith element of which is $\gamma_{i\Delta W}$. Adding the vector y_{t-1} to both sides gives

$$
(6)\quad y_t = Ay^*_t + Zy_{t-1} + \gamma_{\Delta W}\Delta W_t
$$

where $Z = I_5 - A$, with I_5 denoting an identity matrix of order 5. Equations (1), (2), (3), and (6) will be called Model I. They constitute a system of twelve equations in ten endogenous variables (the elements of the vectors y^* and y). There are more equations than unknowns and the question arises as to whether this system is consistent. A sufficient condition for the existence of a solution is given by the following well-known theorem: "A system $CX + b = 0$ of m linear equations in n unknowns is consistent if, and only if, the coefficient matrix C and the augmented matrix $[Cb]$ have the same rank" (Franz Hohn p. 140.) Treating the elements of the vector y as the first five unknowns and the elements of the vector y^* as the next five and writing the twelve equations in the following order—(6), (1), (3), (2)—the 12x10 coefficient matrix, C, of the system is

$$
(7)\quad C = \begin{bmatrix} I_5 & -A \\ 0 & I_5 \\ r' & 0 \\ 0 & r' \end{bmatrix}
$$

The 12x11 augmented matrix $[Cb]$ of the system then is

$$
(8)\quad [Cb] = \begin{bmatrix} I_5 & -A & -(Zy_{t-1} + \gamma_{\Delta W}\Delta W_t) \\ 0 & I_5 & -(B_1X_1 + BX_t)W_t \\ r' & 0 & -W_t \\ 0 & r' & -W_t \end{bmatrix}
$$

According to the theorem, sufficient conditions for these two matrices to have the same rank are also sufficient conditions for the system of equations, Model I, to be consistent. We therefore seek sufficient conditions for the rank of the matrix (8) to have the same rank as the matrix (7). We may form rank-equivalent matrices by elementary row operations. This will be facilitated by partitioning the matrix (8) as follows:

$$
(8')\quad [Cb] = \begin{bmatrix} M_{1-5} \\ M_{6-10} \\ M_{11} \\ M_{12} \end{bmatrix}
$$

where M_{1-5} represents the first five rows of (8), M_{6-10} represents the next five, M_{11} the eleventh, and M_{12} the twelfth row. We now postmultiply r' by M_{1-5} and subtract the result from M_{11} to get

$$
(9)\quad \begin{bmatrix} I_5 & -A & -(Zy_{t-1} + \gamma_{\Delta W}\Delta W_t) \\ 0 & I_5 & -(B_1X_1 + BX_t)W_t \\ 0 & r'A & -W_t + r'(Zy_{t-1} + \gamma_{\Delta W}\Delta W_t) \\ 0 & r' & -W_t \end{bmatrix}
$$

$$(10)\quad \begin{bmatrix} I_5 & -A & -(Zy_{t-1} + \gamma_{\Delta W}\Delta W_t) \\ 0 & I_5 & -(B_1X_1 + BX_t)W_t \\ 0 & 0 & -W_t + r'(Zy_{t-1} + \gamma_{\Delta W}\Delta W_t) + r'A(B_1X_1 + BX_t)W_t \\ 0 & 0 & -W_t + r'(B_1X_1 + BX_t)W_t \end{bmatrix}$$

Partitioning (9) as follows:

$$(9')\quad \begin{bmatrix} N_{1-5} \\ N_{6-10} \\ N_{11} \\ N_{12} \end{bmatrix}$$

where the N_i represent row(s) of (9) and the subscripts are to be interpreted as in (8′), we postmultiply $r'A$ by N_{6-10} and subtract the result from N_{11}, and postmultiply r' by N_{6-10} and subtract the result from N_{12}, to get the matrix that we have denoted as (10). The matrix (10) has the same rank as the matrix (8) and will have the same rank as the matrix (7) if the elements in the eleventh column of the last two rows of this matrix are equal to zero; that is, if

$$(11)\quad \begin{aligned} W_t &= r'(Zy_{t-1} + \gamma_{\Delta W}\Delta W_t) \\ &\quad + r'A(B_1X_1 + BX_t)W_t, \end{aligned}$$

and

$$(12)\quad W_t = r'(B_1X_1 + BX_t)W_t$$

If (11) and (12) are satisfied, our system of twelve equations in ten unknowns will have at least one solution. Considering first the sufficient condition for (12) to be satisfied, we recall that X_1 is a scalar, identically equal to one. Clearly, therefore, (12) will be satisfied if

$$(12a)\quad \begin{cases} r'B_1 = 1 \\ r'B = 0 \end{cases}$$

where 0 is a row vector of six zeros. Conditions (12a) embody the restrictions discussed by Brainard and Tobin (pp. 103, 107) on the structural coefficients of the static equilibrium model. These conditions require constant terms summed over equations (1) to equal unity and coefficients of any interest rate or income so summed to add to zero. If a change in an exogenous variable (interest rate, income) induces an increase in the desired amount of one asset, it induces corresponding decreases (increases) in the desired amounts of some other assets (liabilities). This is the basis of the authors' plea that all interest rates and income should be entered as exogenous variables in the equation of each asset and liability.

Turning next to sufficient conditions for (11) to be satisfied, we first substitute for Z to get:

$$\begin{aligned} W_t &= r'A(B_1X_1 + BX_t)W_t \\ &\quad + r'(I_5 - A)y_{t-1} + r'\gamma_{\Delta W}\Delta W_t \end{aligned}$$

Subtracting $r'y_{t-1}$ from both sides gives:

$$\begin{aligned} W_t - r'y_{t-1} &= r'A[(B_1X_1 + BX_t)W_t \\ &\quad - y_{t-1}] + r'\gamma_{\Delta W}\Delta W_t \end{aligned}$$

In view of (2) we may write:

$$(11')\quad \begin{aligned} \Delta W_t &= r'A[(B_1X_1 + BX_t)W_t \\ &\quad - y_{t-1}] + r'\gamma_{\Delta W}\Delta W_t \end{aligned}$$

Clearly one sufficient set of conditions for (11′) to be satisfied is:

$$(11a)\quad \begin{cases} r'A = 0 \\ r'\gamma_{\Delta W} = 1 \end{cases}$$

where 0 is a row vector of five zeros. Conditions (11a) embody the restrictions discussed by Brainard and Tobin (pp. 106, 108) on the adjustment coefficients of the dynamic system (6). In strict analogy to the case of the structural coefficients, if a gap between desired and actual amounts of a particular asset induces an increase in the actual value of that asset, it induces corresponding de-

creases (increases) in the actual values of other assets (liabilities). Conditions (11a) require the adjustment coefficients summed over a particular gap to add to zero and require the coefficients of the ΔW term to add to unity (as one would surely expect they must since the proportions in which a change in net worth is distributed over assets and liabilities add to 100 percent).

Brainard and Tobin stated that conditions (11a) are necessary and sufficient for satisfaction of (11′). It is clear, however, that if (12a) is satisfied, they are not necessary. An alternative set of conditions exists:

$$\text{(11b)} \qquad \begin{cases} r'A = r' \\ r'\gamma_{\Delta W} = 0 \end{cases}$$

as is easily verified by first substituting (11b) into (11′), noting the restrictions on B_1 and B given in (12a), and noting that by equation (2) $r'y_{t-1} = W_{t-1}$. Conditions (11b) require adjustment coefficients summed over an asset (liability) gap to equal unity (minus one) and require the coefficients of the ΔW term to sum to zero rather than unity. In this case the requirement that the proportions in which a change in wealth is distributed over assets and liabilities add to 100 percent is taken care of by the use of the restriction on the vector B_1 given in conditions (12a).

II

Brainard and Tobin assumed particular values for the elements of the A and B matrices and the $\gamma_{\Delta W}$ vector. However, one might well wish to estimate the system (6) empirically. Several problems arise in this connection. They will be discussed in turn.

We begin by noting that the equation system (6) is deterministic. Treating the scalar W_t and the elements of the vector X_t as fixed in repeated samples we may introduce stochastic elements into it:

$$\text{(6v)} \qquad y_t = Ay_t^* + Zy_{t-1} + \gamma_{\Delta W}\Delta W_t + v_t,$$

where v_t is a 5x1 vector of disturbance terms and we assume that its elements have zero expectation and serial independence and are homoskedastic.

Our complete model now consists of the equations (1), (2), (3), and (6v), which is again a system of twelve equations in ten endogenous variables. We will call this system Model II. It differs from Model I in that the five equations (6) have been replaced by the five equations (6v). Sufficient conditions for consistency of Model II are derived in the same manner used in deriving such conditions for consistency of Model I. They require satisfaction of the following equations:

$$\text{(11v)} \qquad -W_t + r'(Zy_{t-1} + \gamma_{\Delta W}\Delta W_t + v_t) + r'A(B_1X_1 + BX_t)W_t = 0$$

$$\text{(12)} \qquad -W_t + r'(B_1X_1 + BX_t)W_t = 0$$

We have already seen that equation (12) is satisfied by the conditions (12a). To find sufficient conditions for (11v) to be satisfied we proceed as with equation (11). Substituting for Z, subtracting $r'y_{t-1}$ from both sides, and writing ΔW_t for $W_t - r'y_{t-1}$ we get:

$$\text{(11v}'\text{)} \qquad \Delta W_t = r'A[(B_1X_1 + BX_t)W_t - y_{t-1}] + r'(\gamma_{\Delta W}\Delta W_t + v_t)$$

Clearly one sufficient set of conditions for (11v′) to be satisfied is:

$$\text{(11v, a)} \qquad \begin{cases} r'A = 0 \\ r'\gamma_{\Delta W} = 1 \\ r'v_t = 0 \end{cases}$$

where 0 is a row vector of five zeros. However an alternative set of sufficient conditions exists:

$$\text{(11v, b)} \qquad \begin{cases} r'A = r' \\ r'\gamma_{\Delta W} = 0 \\ r'v_t = 0 \end{cases}$$

as is easily verified by first substituting (11v, b) into (11v′), noting the restrictions on B_1 and B given in (12a), and noting that by equation (2) $r'y_{t-1} = W_{t-1}$. Conditions (11v, a) and (11v, b) differ from conditions (11a) and (11b), respectively, in that they further require $r'v_t = 0$. That is, they require the elements of the disturbance vector to sum to zero. The other features of conditions

(11a) and (11b) are preserved in Model II. We now proceed to an interpretation of the difference between conditions (11v, a) and conditions (11v, b).

Substituting (1) into (6v) gives the following system:

$$y_t = A(B_1X_1 + BX_t)W_t + Zy_{t-1} + \gamma_{\Delta W}\Delta W_t + v_t$$

$$(13)\qquad = [AB_1\ AB\ Z\ \gamma_{\Delta W}]\begin{bmatrix} W_t \\ X_tW_t \\ y_{t-1} \\ \Delta W_t \end{bmatrix} + v_t$$

The system of reduced equations for estimating the parameter matrices A and B and vectors B_1 and $\gamma_{\Delta W}$ is given by

$$(13R)\qquad y_t = [\gamma_1\quad \Gamma_{2,7}\quad \Gamma_{8,12}\quad \gamma_{\Delta W}]\begin{bmatrix} W_t \\ X_tW_t \\ y_{t-1} \\ \Delta W_t \end{bmatrix} + v_t$$

where γ_1 is the 5x1 vector of reduced form coefficients of W_t, $\Gamma_{2,7}$ is the 5x6 matrix of reduced form coefficients of the exogenous variables, $\Gamma_{8,12}$ is the 5x5 matrix of reduced form coefficients of the lagged endogenous variables, and $\gamma_{\Delta W}$ is the 5x1 vector of coefficients of the ΔW term. But unique estimates of γ_1, $\Gamma_{8,12}$ and $\gamma_{\Delta W}$ cannot be obtained since an exact linear relation holds between W_t, ΔW_t and the five elements of y_{t-1}:

$$(2')\qquad W_t = \Delta W_t + r'y_{t-1}$$

The problem can be made explicit. We may substitute (2′) into (13R) to eliminate W_t from the latter.[2] Upon collecting terms we get

$$(13Ra)\qquad y_t = [\Gamma_{2,7}, \Gamma_{8,12} + \gamma_1 r', \gamma_1 + \gamma_{\Delta W}] \cdot \begin{bmatrix} X_tW_t \\ y_{t-1} \\ \Delta W_t \end{bmatrix} + v_t$$

It is quite clear that the individual components of the two sums $\Gamma_{8,12}+\gamma_1 r'$ and $\gamma_1+\gamma_{\Delta W}$ are not identifiable. However, if we are willing to adopt the expedient of setting all components of the vector γ_1 equal to zero, the elements of $\Gamma_{8,12}$ and $\gamma_{\Delta W}$ are identifiable.

Instead of using (2′) to eliminate W from (13R) we might just as easily have used it to eliminate ΔW. In that case we would get

$$(13Rb)\qquad y_t = [\Gamma_{2,7}, \Gamma_{8,12} - \gamma_{\Delta W}, \gamma_1 + \gamma_{\Delta W}] \cdot \begin{bmatrix} X_tW_t \\ y_{t-1} \\ W_t \end{bmatrix} + v_t$$

We cannot separately identify the components of $\Gamma_{8,12}-\gamma_{\Delta W}$ and of $\gamma_1+\gamma_{\Delta W}$ but if we set all the elements of the vector $\gamma_{\Delta W}$ equal to zero, the elements of $\Gamma_{8,12}$ and of γ_1 are identifiable.[3]

To assume that the elements of γ_1 are all equal to zero and all the elements of X_t are not constant over time implies that W_t does not enter equation system (1) as an isolated variable in one of the linear terms. This in turn implies there is no functional dependence of the elements of the vector of equilibrium values of assets and liabilities, y_t^*, on net worth, apart from a scale factor. The first difference of net worth continues to enter the adjustment equations (6v), however, and does so without a coefficient of lagged adjustment. Therefore, to omit the variable W_t from system (1), while retaining ΔW_t in (6v) is to assume that the endogenous variables adjust instantaneously to a change in net worth.

On the other hand, to assume that the elements of $\gamma_{\Delta W}$ are all equal to zero is to omit the variable ΔW_t from equation system (6). If this is done it would appear that changes in assets and liabilities do not depend on net worth. However, since W_t continues to enter equations (1) which determine y^*, and since changes in assets and liabilities do depend on y^*, they also de-

[2] Of course we may also substitute (2′) into the term X_tW_t but no purpose is served by doing so, since that term has nothing to do with the problem under discussion.

[3] Of course we can also achieve identifiability in a large number of other ways. Among other possibilities we might set any column of $\Gamma_{8,12}$ equal to zero.

pend on net worth. Furthermore, since these changes represent only partial adjustment to changes in y^*, they represent only partial (or lagged) adjustment to changes in net worth.

It follows that if W_t is dropped from the reduced equation system (13R) and ΔW_t enters these equations, instantaneous adjustment to a change in net worth is assumed and the conditions (11v,a) apply. If ΔW_t is omitted, W_t enters these equations, lagged adjustment to a change in net worth is assumed, all elements of the vector $\gamma_{\Delta W}$ are zero, and the conditions (11v, b) apply.

We must next show how we derive estimates of the elements of A, B, and B_1 from estimates of the parameters of the system (13R). Comparing (13) with (13R) it is seen that estimates of A are easily obtained. Since $Z = I_5 - A$, $A = I_5 - \Gamma_{8,12}$, and

(14) $$\hat{A} = I_5 - \hat{\Gamma}_{8,12}$$

where hats indicate estimated values. Careful comparison of the two systems also shows that

$$\Gamma_{2,7} = AB$$
$$\gamma_1 = AB_1$$

Consider the case in which W is dropped from the regression equations. The elements of γ_1 are all equal to zero and estimates of the elements of $\gamma_{\Delta W}$ are given directly by the regression coefficients of ΔW. Estimates of A are obtained using (14). Denote this estimated matrix of adjustment coefficients as $\hat{A}^a$, and denote the matrix of estimates of $\Gamma_{2,7}$ as $\hat{\Gamma}^a_{2,7}$. Then

(15a) $$\hat{\Gamma}^a_{2,7} = A^a\hat{B}^a$$

where $\hat{B}^a$ is an estimate of the matrix B and the superscript a indicates that W has been omitted from the relations. It would seem that the solution for $\hat{B}^a$ is straightforward until one recalls the conditions (11a). They imply that the matrix A is singular. Assuming the estimate of A, $\hat{A}^a$, satisfies conditions (11a) it also is singular and no unique solution would seem to exist for $\hat{B}^a$. However we may add conditions (12a) to the system (15a) as follows:

(15a′) $$\begin{bmatrix} \hat{\Gamma}^a_{2,7} \\ 0 \end{bmatrix} = \begin{bmatrix} A^a \\ r' \end{bmatrix} \hat{B}^a$$

where 0 is a 1x6 row vector of zeros. Deleting all but one column from the left-hand matrix and all but the same column from $\hat{B}^a$, we get:

(15a″) $$\begin{bmatrix} \hat{\Gamma}^a_j \\ 0 \end{bmatrix} = \begin{bmatrix} A^a \\ r' \end{bmatrix} [\hat{B}^a_j]$$

where the left-hand matrix denotes the jth column of the left-hand matrix of (15a′) and $\hat{B}^a_j$ denotes the jth column of $\hat{B}^a$. Then (15a″) is a system of six equations in five unknowns. We can find the requirement for this system to be consistent in the same way as we proceeded in the case of the system (8). The requirement is that $r'\hat{\Gamma}^a_j = 0$. If our estimate of Γ_j, $\hat{\Gamma}^a_j$, satisfies this condition, then by removing a dependent row of elements with row index i from $\hat{A}^a$ and the element with the same row index from $\hat{\Gamma}^a_j$, the system is transformed into one of five independent equations which may be solved for the five unknowns $\hat{B}^a_j$.[4] The process may be repeated five times until a solution for the entire matrix $\hat{B}^a$ has been obtained.

If ΔW has been omitted from the equations, the elements of $\gamma_{\Delta W}$ are all equal to zero and estimates of the elements of A, $\hat{A}^b$ say, are obtained using (14). If $\hat{\gamma}_1$ is the vector of estimates of the elements of γ_1, and $\hat{\Gamma}^b_{2,7}$ is the matrix of estimates of the elements of $\Gamma_{2,7}$, then we wish to solve the system

(15b) $$\hat{\gamma}_1\hat{\Gamma}^b_{2,7} = \hat{A}^b[\hat{B}_1\hat{B}^b]$$

for $[\hat{B}_1\hat{B}^b]$, an estimate of $[B_1B]$. Since conditions (11b) do not imply the singularity of A, the assumption that $\hat{A}^b$ satisfies these conditions does not present the same difficulty as in the preceding paragraph, and the solution process is straightforward. However, the solution should be consistent with condi-

[4] If the rank of $\hat{A}^a$ is four (or more generally, if the rank is one less than the order of $\hat{A}^a$), consistency of (15a″) guarantees uniqueness of the solution, $\hat{B}^a_j$, regardless of the choice of the dependent row index i. I am indebted to M. J. Ringo for spotting an error on this point in an earlier draft.

tions (12a). To investigate the requirements for such consistency we form

$$\text{(15b}')\qquad \begin{bmatrix} \hat{\gamma}_1 & \hat{\Gamma}^b_{2,7} \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} A^b \\ r' \end{bmatrix} [\hat{B}_1 \hat{B}^b]$$

and proceed to analyze it in the manner that system (8) was analyzed. The requirements for consistency turn out to be

$$r'\hat{\Gamma}^b_{2,7} = 0$$

$$r'\hat{\gamma}_1 = 1$$

Hence our estimates of $\Gamma_{2,7}$ and of γ_1 must satisfy these conditions, if the solution for $[\hat{B}_1\hat{B}^b]$ is to satisfy both conditions (11b) and (12a).

Finally, the question arises as to whether ordinary least squares estimates of $\Gamma_{8,12}$ yield estimates of A (by the transformation (14)) which satisfy (11a) or (11b) as may be appropriate, or whether some constrained estimation technique is necessary. The answer is that so long as all lagged endogenous variables appear in all equations of (13R), the ordinary least squares estimator of $\Gamma_{8,12}$ can be transformed into an estimate of A which satisfies the restrictions. We first prove the proposition for the case in which W is dropped. Our data are constrained by (2′) which may be rewritten

$$r'y_t = r'y_{t-1} + \Delta W_t$$

or

$$r'y_t = [0 \quad r' \quad 1] \begin{bmatrix} X_t W_t \\ y_{t-1} \\ \Delta W_t \end{bmatrix}$$

where 0 is a 1×6 vector of zeros, or

$$\text{(16a)}\qquad r'y_t = q'f_a$$

where $q' = [0 \quad r' \quad 1]$ and $f_a = \begin{bmatrix} X_t W_t \\ y_{t-1} \\ \Delta W_t \end{bmatrix}$

Transposing (16a) and writing it repeatedly for all T observations yields

$$\text{(16a}')\qquad Yr = F_a q$$

where Y and F_a are the Tx5 and Tx12 data matrices on the variables. Assuming the elements of the vector γ_1 are all equal to zero, the ordinary least squares estimates of the parameters of equations (13Ra) are given by

$$\text{(17a)}\qquad \hat{\Gamma}^{a\prime} = (F_a'F_a)^{-1}F_a'Y$$

where Γ^a is the parameter matrix $[\Gamma_{2,7}, \Gamma_{8,12}, \gamma_{\Delta W}]$. Postmultiplying (17a) by r, and substituting Yr from (16a′) yields

$$\hat{\Gamma}^{a\prime}r = (F_a'F_a)^{-1}F_a'F_a q$$

or

$$\hat{\Gamma}^{a\prime}r = q$$

Transposing, one gets

$$r'\hat{\Gamma}^a = q'$$

By comparing system (13) with (13Ra) one sees that $\hat{\Gamma}^a$ is an estimate of the parameter matrix $[AB \; Z \; \gamma_{\Delta W}]$. Making this substitution, and also substituting the definition of q', one gets

$$\text{(18a}')\qquad r'\hat{Z} = r'$$

$$\text{(18a}'')\qquad r'A\hat{B} = 0$$

$$\text{(18a}''')\qquad r'\gamma_{\Delta W} = 1$$

Since $\hat{Z} = I_5 - \hat{A}^a$, substituting into (18a′) gives

$$r'(I_5 - A^a) = r'I_5$$

or $r'\hat{A}^a = 0$. Since we also have (18a‴), we have shown that our estimates satisfy conditions (11a).

Next consider the case in which ΔW is omitted. We rewrite (2′) as

$$r'y_t = [0 \quad 1 \quad 0] \begin{bmatrix} y_{t-1} \\ W_t \\ X_t W_t \end{bmatrix}$$

where the 0's represent 1×5 and 1×6 vectors of zeros, or

$$\text{(16b)}\qquad r'y_t = p'f_b$$

where $p' = [0 \quad 1 \quad 0]$ and $f_b = \begin{bmatrix} y_{t-1} \\ W_t \\ X_t W_t \end{bmatrix}$

Transposing (16b) and writing it repeatedly for all T observations yields

$$Yr = F_b q \tag{16b'}$$

where Y and F_b are the Tx5 and Tx12 data matrices on the variables. Assuming the elements of the vector $\gamma_{\Delta W}$ are all equal to zero, the ordinary least squares estimates of the parameters of equation (13Rb) are given by

$$\hat{\Gamma}^{b'} = (F_b' F_b)^{-1} F_b' Y \tag{17b}$$

where Γ^b is the parameter matrix $[\Gamma_{2,7}, \Gamma_{8,12}, \gamma_1]$. Postmultiplying (17b) by r, and substituting Yr from (16b′) yields

$$\hat{\Gamma}^{b'} r = (F_b' F_b)^{-1} F_b' F_b p$$

or $\hat{\Gamma}^{b'} r = p'$. Transposing, one gets $r'\hat{\Gamma}^b = p'$. By comparing equations (13Rb) with (13) one sees that $\hat{\Gamma}^b$ is an estimate of the parameter matrix $[AB_1 \; AB \; Z]$. Making this substitution, and also substituting the definition of p', one gets

$$r'\hat{Z} = r' \tag{18b'}$$

$$r' A\hat{B}_1 = 1 \tag{18b''}$$

$$r' A\hat{B} = 0 \tag{18b'''}$$

Since $\hat{Z} = I_5 - \hat{A}^b$, substituting into (18b′) gives $r'(I_5 - \hat{A}^b) = 0$ or $r'\hat{A}^b = r'$, which is the first condition in (11b). The second condition, $r'\gamma_{\Delta W} = 0$ is satisfied by assumption. Substituting $r'\hat{A}^b = r'$ into (18b″) and (18b‴) gives conditions (12a) which, hence, are also satisfied by the ordinary least squares estimators.

III

The model of Brainard and Tobin consists of identities which imply restrictions on the coefficients of their static and dynamic behavioral equations. While Brainard and Tobin recognized these restrictions, indeed emphasized them, they did not derive them formally. Such a derivation has been accomplished in Section I of this essay. In Section II we showed how estimates of the structural coefficients of the model which satisfy the restrictions derived in Section I could be derived from ordinary (unconstrained) least squares estimates of a reduced system of equations.

REFERENCES

W. C. Brainard and J. Tobin, "Pitfalls in Financial Model Building," *Amer. Econ. Rev. Proc.*, May 1968, *58*, 99–122.

F. E. Hohn, *Elementary Matrix Algebra*, New York 1964.

Clothing Exemptions and Sales Tax Regressivity: Note

By David G. Davies*

In a recent article in this *Review*, Jeffrey Schaefer reports an important discovery that runs counter to currently accepted views. He finds that clothing exemptions in the New Jersey tax law actually "reduce sales tax progressivity."[1] It would be of great value if Schaefer's results could be generalized to a wider geographical area. Other state legislatures considering adopting or revising the sales tax, for example, could be spared the mistake of exempting clothing in the belief that such an exemption pushes the sales tax toward progressivity if, in fact, the opposite is true.

This comment extends Schaefer's analysis to the whole of the United States. We derive progressivity-regressivity indexes for a sales tax which excludes clothing from taxation for 1) all urban areas in the United States, and 2) all urban and rural areas combined. The empirical information for these areas comes from the most recent survey of consumer expenditures and income for the United States by the Bureau of Labor Statistics (1964 and 1966).

Following Schaefer's definitions, a sales tax is considered to be progressive if the effective rate of taxation increases as the ability to pay increases. If the effective tax rate declines as the ability to pay increases, the tax is regressive; and if the rate remains approximately constant as ability to pay changes, the tax is proportional.

An effective way to test whether a sales tax is progressive or not is to derive the elasticity of the tax base with respect to the measure of ability to pay. We can use Schaefer's regression equation to derive the progressivity-regressivity index of alternative sales tax bases:

$$\sum_{j=1}^{n} w_j \log \bar{t}_{ij} = N \log a_i + b_{ik} \sum_{j=1}^{n} w_j \log \bar{y}_{jk} + \sum_{j=1}^{n} w_j \log e_{ijk} \tag{1}$$

where

$\bar{t}_{ij}$ = mean ith sales tax base for the jth income class,

$\bar{y}_{jk}$ = mean income of the jth income class under the kth income concept,

b_{ik} = the progressivity-regressivity index for the ith sales tax base under the kth income concept,

a_i = the constant, and

e_{ijk} = the error term for the ith sales tax base for the jth income class under the kth income concept.

If the regression coefficient, b, is greater than one, that particular sales tax base is progressive. If b is approximately equal to one, the tax is proportional; and it is regressive if b is less than one.

Although annual measured money income is a fairly generally accepted criterion of ability to pay, we shall also use Irving Fisher's notion of income and Harold Somers' and Joseph Launie's concept of net resources as alternative measures of ability to pay in deriving the progressivity-regressivity indexes for alternative tax bases. Conceptually, Fisher income is equal to consumption,[2] and net resources are defined as annual net income plus net worth.

* Professor of economics, Duke University. I wish to express my appreciation to Professor Jay Salkin and Mr. Stanley Warner for advice and research assistance. I would also like to thank an anonymous referee for comments which improved the paper considerably.

[1] Although there has been no published empirical study of the equity effect of taxing clothing other than Schaefer's contribution, the currently accepted view is that including clothing in a sales tax base helps to make the levy regressive. See, for example, Alfred D. Buehler (p. 229), Frank Greenway, Paul Hastings, and John Smale (pp. 434–36), Niel Jacoby (p. 183), and American Federation of Labor and Congress of Industrial Organization (p. 77).

[2] Fisher reasoned that consumption is the destruction of utility and comes closest to measuring real income. See Irving Fisher.

TABLE 1—PROGRESSIVITY-REGRESSIVITY INDEXES OF DIFFERENT SALES TAX BASES USING ANNUAL INCOME, FISHER INCOME, AND NET RESOURCES AS ALTERNATIVE MEASURES OF ABILITY TO PAY[a]

	Urban Population			Urban Plus Rural Population		
	Annual Income	Fisher Income	Net Resources	Annual Income	Fisher Income	Net Resources
T_1 = Taxable commodities (normal tax base in food taxing states)	.87 (.04)	1.08 (.02)	1.10 (.14)	.79 (.05)	1.03 (.02)	1.04 (.14)
$T_2 = T_1$ minus clothing expenditures	.84 (.05)	1.04 (.02)	1.05 (.14)	.77 (.05)	1.00 (.02)	1.00 (.14)
T_3 = Taxable commodities (normal tax base in states not taxing food)	.97 (.04)	1.21 (.02)	1.24 (.14)	.88 (.04)	1.14 (.01)	1.16 (.13)
$T_4 = T_3$ minus clothing expenditures	.94 (.05)	1.17 (.02)	1.20 (.14)	.86 (.05)	1.11 (.01)	1.13 (.14)
T_5 = Clothing expenditures	1.07 (.03)	1.32 (.04)	1.38 (.13)	.95 (.04)	1.22 (.03)	1.27 (.12)
T_6 = Expenditures on children's (under 18) clothing	1.27 (.07)	1.58 (.05)	1.61 (.21)	1.04 (.08)	1.36 (.03)	1.37 (.18)

[a] Standard errors are in parentheses.

The coefficients in Table 1 reveal that the exemption of clothing always pushes the sales tax toward regressivity, regardless of whether or not a state includes or excludes food from the tax base,[3] whether or not annual measured money income, Fisher income, or net resources are used to calculate the equity coefficient, and whether or not we consider the urban population or the entire urban-rural population of the United States.

A tax base of clothing per se would yield a progressive tax for urban residents in the United States if net annual income, Fisher income, or net resources are used as the measures of ability to pay. A tax on rural plus urban expenditures for clothing is roughly proportional on net income, but progressive for Fisher income and net resources.

We also find from the coefficients in Table 1 that the progressivity-regressivity indexes of spending for children's clothing exceeds the indexes of total clothing expenditures. Moreover, a tax on children's clothing is progressive for all geographical areas and income concepts with the exception of annual income for the combined urban-rural total where the levy is proportional. For equity purposes, exempting children's clothing from taxation, as Connecticut has done, would appear to be worse than exempting total clothing.[4]

In conclusion, it seems clear that Schaefer's important findings about New Jersey can be extended to cover both the urban and entire rural-urban population of the United States. All states considering adoption or changes in sales tax laws should be aware that exempting clothing from taxation not only erodes the tax base, but makes the levy less progressive or more regressive.

[3] Currently thirty states include food while fifteen exclude it from the tax base.

[4] The relatively high elasticity of children's clothing may be explained by the positive correlation between income and number of children.

REFERENCES

A. D. Buehler, *General Sales Taxation*, New York 1932.

I. Fisher and H. W. Fisher, *Constructive Income Taxation*, New York 1942.

F. Greenway, P. Hastings, and J. Smale, "Consumption Taxes" in *Public Finance*, New York 1959.

N. Jacoby, *Retail Sales Taxation,* Chicago 1938.

J. M. Schaefer, "Clothing Exemptions and Sales Tax Regressivity," *Amer. Econ. Rev.*, Sept. 1969, *59*, 596–99.

H. M. Somers assisted by J. J. Launie, *The Sales Tax,* Sacramento 1964.

American Federation of Labor and Congress of Industrial Organization, *State and Local Taxes,* Washington 1958.

U.S. Bureau of Labor Statistics, *Consumer Expenditures and Income, Urban United States, 1960–61,* Washington 1964.

———, *Consumer Expenditures and Income, Urban and Rural United States, 1960–61,* Washington 1966.

An Economic Theory of the Second Moments of Disturbances of Behavioral Equations

By HENRI THEIL*

Applied econometricians have known for a long time that the smooth curves in economics texts which pretend to be the geometric representations of behavioral equations present an over-simplified picture of the real world. It almost never occurs that data fit a simple theoretical relation exactly, and an explicit method of interpreting and handling the deviations from such relations is therefore needed. In the early days of econometrics it was frequently assumed that observational errors were the only or at least the main cause of the occurrence of deviations, but this idea was later rejected on the ground that even when the data are perfect, deviations will continue to be found because there are very few economic agents who react exactly according to some simple mathematical law. The "neglected factors" entered into the picture; their combined influence came to be regarded as a random disturbance with zero mean and unknown variance.

It is important to realize that this disturbance variance is as much an unknown parameter of the model as the coefficients of the explanatory variables are. Typically, however, the economic theorist concentrates his efforts exclusively on these variables and their coefficients. He assumes that the decision maker whose behavior is described by the equation maximizes or minimizes a criterion function:

$$(1) \qquad f(x) = f(x_1, \ldots, x_n)$$

where $x=[x_1 \ldots x_n]'$ is a vector of decision variables, and that the extremum is sought subject to a set of q constraints:

$$(2) \qquad g_h(x) = 0 \qquad h = 1, \ldots, q$$

Under appropriate conditions on the form of the functions (1) and (2), he will obtain as the solution a vector $\bar{x}$ which expresses each decision variable uniquely in terms of the parameters that determine the functions $f(\ )$, $g_1(\ ), \ldots, g_q(\ )$. Consumer demand theory is a classical example; then (1) is the utility function and (2) is the budget constraint, and the decision variables $x_1, \ldots, x_n$ are the quantities bought which are expressed by the demand functions (the behavioral equations) in terms of income and prices (the $n+1$ parameters of the budget constraint).

The disturbing thing about this approach is that it leaves no room for a random disturbance. It is undoubtedly true that there are factors which are neglected by the approach, so that discrepancies between the behavior predicted by the equation(s) and the observed behavior must be expected, but it is equally true that the economic theory in this form has nothing to say about the variance of such discrepancies. This problem has gained in importance in the last few decades since the introduction of simultaneous equation systems and "seemingly unrelated regressions," which require consideration of the disturbance variances of several equations and also of their contemporaneous covariances. When there are L such equations and when the jth contains N_j unknown coefficients, their total number is

$$(3) \qquad N_1 + N_2 + \ldots + N_L$$

and the total number of variances and contemporaneous covariances (taking account of the symmetry of their matrix) is

$$(4) \qquad (\tfrac{1}{2})L(L+1)$$

When L is moderately large, the number (4) may be of the same order of magnitude as the number (3) or even larger.

Economic theory in its conventional form focuses on the expectational part of behavioral equations by making more or less

* University of Chicago.

explicit statements on the coefficients whose number is given in (3). The question arises whether it is possible to extend this theory in such a way that it can also be brought to bear on the second moments whose number is given in (4). The objective of this article is to show that, if a set of behavioral equations refers to one single decision maker (as in the consumer demand case discussed following (2)), a plausible argument can be made for the proposition that the Hessian matrix[1] of the criterion function is the crucial determinant of the covariance matrix of the disturbances.[2] Basically, the idea in the case of one single decision variable is that, if the second-order derivative at the point of the extremum is close to zero, the loss incurred by a decision which deviates from the optimum to a moderate extent is very small, and the disturbance of the corresponding behavioral equation may then be expected to have a large variance.

I. The Unconstrained Case

We start with the case in which there are no constraints, q of (2) being zero, and assume that $f(\ )$ of (1) is to be minimized. In many economic theories, this is a smooth and well-behaved function and we shall confine ourselves to this type by assuming that it can be approximated to a sufficient degree of accuracy by a quadratic function around the minimum:

$$f(x) = a'x + (\tfrac{1}{2})x'Ax \tag{5}$$

where the constant term is disregarded because it is irrelevant from the viewpoint of minimization. The matrix A is the (symmetric) Hessian matrix of $f(\ )$ evaluated at the point of the minimum; it will be assumed to be positive definite. The decision vector which minimizes $f(\ )$ is then

$$\hat{x} = -A^{-1}a \tag{6}$$

Consider any decision x which differs from $\hat{x}$. The associated value $f(x)$ of the criterion function will exceed $f(\hat{x})$ and the excess is the following quadratic form:

$$\begin{aligned} f(x) - f(\hat{x}) &= \tfrac{1}{2}(x-\hat{x})'A(x-\hat{x}) \\ &= \tfrac{1}{2}\xi'A\xi \end{aligned} \tag{7}$$

where

$$\xi = x - \hat{x} = x + A^{-1}a \tag{8}$$

is the decision discrepancy vector. Given that A is symmetric and positive definite, there exists a nonsingular matrix P such that

$$P'P = A \tag{9}$$

It follows from (7) that the excess can be written as

$$f(x) - f(\hat{x}) = \tfrac{1}{2}\xi'P'P\xi = \tfrac{1}{2}\eta'\eta \tag{10}$$

where

$$\begin{aligned} \eta = P\xi &= Px + PA^{-1}a \\ &= Px + P(P'P)^{-1}a \\ &= Px + (P')^{-1}a \end{aligned} \tag{11}$$

The second and third steps in (11) are based on (8) and (9), respectively.

Equation (10) shows that minimizing $f(\ )$ is equivalent to the minimization of the sum of squares of the elements of η. This is the simplest possible reduction of the extremum problem, and we shall therefore refer to the η elements as the ***elementary decision variables.*** If the decision maker definitely wants to minimize $f(\ )$, he should put each of these variables equal to zero. If he behaves differently, the loss which he incurs—the excess of $f(\ )$ over the minimum value—is equal to one-half of the squared length of η.

To introduce the stochastic element we suppose that the vector a of the linear part of $f(\ )$ in (5) fluctuates randomly, while the matrix A of the quadratic part remains fixed as before.[3] It follows from (6) that the opti-

[1] The matrix of second-order derivatives.

[2] The reader who is statistically oriented will recognize the formal similarity with the asymptotic covariance matrix of maximum-likelihood estimators, which is related to the Hessian matrix of the logarithmic likelihood function.

[3] The simplest way to visualize the case of a stochastic vector a and a nonstochastic matrix A is by interpreting $f(\ )$ as a cost function. Then all marginal cost functions are subject to additive random shocks. If the matrix A were also random, the marginal cost functions would be subject to both additive and multiplicative random shocks, which is much more difficult to handle.

mal vector $\hat{x}$ then becomes a linear function of random variables, so that it is also random. The vector $\hat{x}$ is then no longer feasible as a decision vector. Furthermore, when x is the (nonrandom) decision which is actually made, η of (11) is equal to the sum of the random vector $(P')^{-1}a$ and the nonrandom vector Px. In other words, the decision maker has no longer full control over his elementary decision variables. Part of each such variable is controlled, but the rest is a linear function of random variables.

The minimization of $f(\ )$ no longer suffices as a criterion of this stochastic situation. The criterion to be developed here rests on the assumption that the decision maker is interested in a simple mechanism to control the situation. This seems reasonably realistic, given that the decision makers involved are typically consumers or entrepreneurs who have little inclination toward formal analysis. Specifically, consider the following:

1) Given that η is random, the decision maker no longer controls this vector, but he can at least adjust the decision vector x over time so that he controls the *distribution* of η. It follows from (10) that the loss associated with any decision is equal to $\frac{1}{2}\eta'\eta$, which is a random variable. We assume that the decision maker is interested in controlling its average value, $\frac{1}{2}\mathrm{E}(\eta'\eta)$, and this implies that the only feature of the distribution of η which is relevant is the matrix of its second-order moments.

2) Since the criterion $\frac{1}{2}\mathrm{E}(\eta'\eta)$ concerns only the squares of the elementary decision variables, not their products, there is no gain in designing a particular cross-moment pattern. The simplest solution is then to arrange a diagonal moment matrix $\mathrm{E}(\eta\eta')$.

3) For each of the elementary decision variables, the loss associated with a nonzero value is equal to one-half of its square. There is therefore no reason to let any such variable have a larger mean square than any other, which leads to the specification of equal diagonal elements in the moment matrix $\mathrm{E}(\eta\eta')$. On combining this with the conclusion reached under (2) we obtain

$$\mathrm{E}(\eta\eta') = \sigma^2 I \tag{12}$$

where σ^2 is a measure for the loss which the decision maker is willing to incur on the average. The criterion mentioned under (1), $\frac{1}{2}\mathrm{E}(\eta'\eta)$, is then σ^2 multiplied by one-half the number of decision variables.

It follows from (8) and (11) that the difference between the actual decision x and the theoretically optimal[4] decision $\hat{x}$ is $\xi = P^{-1}\eta$. This difference is the disturbance vector of the equation system which describes the decision maker's behavior with respect to the variables which he controls. Given the specification (12), we obtain the following moment matrix for this disturbance vector:

$$\begin{aligned} \mathcal{V}(x - \hat{x}) &= P^{-1}\mathrm{E}(\eta\eta')(P')^{-1} \\ &= \sigma^2(P'P)^{-1} = \sigma^2 A^{-1} \end{aligned} \tag{13}$$

which thus turns out to be a positive multiple of the inverse of the Hessian matrix of the criterion function.

When there is only one decision variable, this result implies a disturbance variance which is inversely proportional to the second-order derivative d^2f/dx^2 evaluated at the point of the minimum.[5] As stated at the end of the introduction, this simply means that the disturbance variance is larger (smaller) when the loss associated with a given discrepancy $x-\hat{x}$ is smaller (larger). When we have two decision variables, the moment matrix (13) becomes

$$\begin{aligned} &\sigma^2 \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}^{-1} \\ &= \frac{\sigma^2}{a_{11}a_{22} - a_{12}^2} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{12} & a_{11} \end{bmatrix} \end{aligned} \tag{14}$$

This indicates that, when the two decision variables do not interact in the criterion function (5) in the sense that a_{12} vanishes, the associated disturbances have a zero

[4] Theoretically optimal in the sense of optimal under nonstochastic conditions.

[5] When the criterion function is not quadratic, this derivative is a function of x and its value at $\hat{x}$ is random when $\hat{x}$ is random. One may then decide to take the expectation to obtain an approximate value. Note further that the word "variance" used here in the text is applicable only when the disturbances have zero mean. This condition is usually satisfied due to the presence of a contant term in the equation.

cross-moment. If they do interact, the cross-moment has a sign opposite to that of a_{12}. If they interact to a sufficient degree, a_{12}^2 being close to $a_{11}a_{22}$, the moments are all large in absolute value and the moment matrix is close to singularity. These results are intuitively plausible.

We conclude by noting that the model developed above treats the decision which is actually made (x) as nonstochastic and the theoretically optimal decision and its determining factors ($\hat{x}$ and a) as random variables, whereas it is customary in the standard linear regression model to treat the dependent variable (our decision variable) as random and the explanatory variables (the determining factors) as taking fixed values. This difference is not essential, however, because the conditional distribution of $x-\hat{x}$ given $\hat{x}$ has the same moment matrix as the conditional distribution given x.

II. The Constrained Case

Next assume that there are $q>0$ constraints (2) subject to which the function (5) is to be minimized. We shall linearize these constraints in the same way as the criterion function is quadratized:

$$Bx = b \tag{15}$$

where the elements of B are to be interpreted as first-order partial derivatives of the functions $g_1(\ ), \ldots, g_q(\ )$ defined in (2), evaluated at the point of the constrained minimum. It is assumed that B has full row rank (i.e., rank q), so that the possibility of inconsistent or linearly dependent constraints is excluded.

One way of solving this more general problem is by using the constraints (15) to eliminate q of the x's and then proceeding along the lines of the previous section. However, a more elegant extension can be formulated on the basis of the Lagrangian function:

$$\begin{aligned} F(x, \lambda) &= f(x) + \lambda'(Bx - b) \\ &= [a' \ \ -b'] \begin{bmatrix} x \\ \lambda \end{bmatrix} \\ &\quad + \tfrac{1}{2}[x' \ \ \lambda'] \begin{bmatrix} A & B' \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} \end{aligned} \tag{16}$$

The new decision vector is $[x' \ \lambda']'$ rather than x, the coefficient vector of the linear part is now $[a' \ \ -b']$,[6] and the original Hessian matrix is bordered by q rows and q columns. It is readily seen that the extension of the moment matrix result (13) to the present case amounts to σ^2 times the inverse of the bordered Hessian matrix, and that the leading sub-matrix (corresponding to A in the bordered Hessian before inversion) of this inverse gives the moment matrix of the disturbance vector corresponding to x. Application of the standard rules for partitioned inversion gives:

$$\mathcal{V}(x - \hat{x}) = \sigma^2 [A^{-1} - A^{-1}B'(BA^{-1}B')^{-1}BA^{-1}] \tag{17}$$

The matrix (17) is zero in the special case of a square nonsingular B. This is as it should be, because the decision vector can then be solved from the constraint (15). Note further that premultiplication of the moment matrix (17) by B gives a zero matrix, which is also in agreement with (15).

III. Application to Consumer Demand Theory

Consider a consumer whose objective is to maximize a utility function $u(x)$ subject to the budget constraint $p'x=m$, where p and x are n-element vectors of prices and quantities, respectively, and m is total expenditure (or income). It is assumed that p and m are fixed and given from the consumer's point of view. This amounts to the following specification of the more general model described above: $f(\ )=-u(\ )$, $B=p'$, $b=m$. Write U for the Hessian matrix of $u(\ )$ in the constrained maximum, so that A is specified as $-U$. The moment matrix (17) then becomes

$$-\sigma^2\left(U^{-1} - \frac{1}{p'U^{-1}p}\, U^{-1}pp'U^{-1}\right) \tag{18}$$

which is a positive semi-definite matrix of order $n\times n$ and rank $n-1$ when U is negative definite. The singularity of the moment matrix is due to the budget constraint $p'x=m$.

[6] The form of this coefficient vector indicates that the vector b can be allowed to fluctuate randomly over time, just as a.

The matrix (18) is precisely the same as the moment matrix which I derived earlier on the basis of *ad hoc* considerations.[7] These considerations do add to the plausibility of the general model, however, and it is therefore worthwhile to summarize them here briefly without going into mathematical details:

(1) When equating $-f(x)$ of (5) to $u(x)$, we obtain

$$-a_i + \tfrac{1}{2} \sum_j u_{ij} x_j$$

for the marginal utility of the ith commodity, a_i and u_{ij} being the ith element of a and the (i, j)th element of U, respectively. The assumption of a random vector a implies that the marginal utilities are subject to additive random shocks. Assume that they have zero expectation. The main problem, to be considered under (2), is the specification of the $n \times n$ covariance matrix of these shocks.

(2) Consider the case $u_{ij}=0$, so that the ith and jth marginal utilities do not depend on x_j and x_i, respectively. This indicates that the two commodities satisfy unrelated wants, so that it seems reasonable to assume that the random shocks of their marginal utilities are uncorrelated. Next consider the case of a *negative* u_{ij}. The ith marginal utility then decreases with x_j, which may be interpreted in the sense that the two commodities satisfy similar wants. When the marginal utilities are subject to random fluctuations, those corresponding to similar wants will usually be of the same sign and will therefore be *positively* correlated. Third, consider a commodity with a large negative u_{ii}, so that its marginal utility is unstable in the sense of being sensitive to small changes in q_i. It is then plausible that it is also unstable in the sense that the random shocks of its marginal utility have a large variance. All three cases ($u_{ij}=0$, $u_{ij}<0$, $-u_{ii}$ large) are covered when it is assumed that the covariance matrix of the random shocks is a negative multiple of the Hessian matrix U, and this leads directly to the matrix (18).

[7] See Theil, pp. 228–33.

REFERENCE

H. Theil, *Economics and Information Theory*, Amsterdam; Chicago 1967, pp. 228–33.

A Neglected Social Cost of a Voluntary Military

By THOMAS E. BORCHERDING*

It would appear without exception that economists believe that a voluntary military is preferable to conscription.[1] It is my purpose to demonstrate that this institutional preference is questionable on purely a priori grounds. A *potentially* important welfare cost may arise under voluntarism from the monopsonistic behavior of the defense establishment[2] as a purchaser of enlisted personnel. To analyze this possibility it will be necessary to develop a terse and simple model of choice in the "market" for enlisted personnel and to apply it to the institutions of conscription and voluntarism.

Assume that the process of political exchange is efficient to a degree that the military's demand function for enlisted men is a close approximation to the social marginal value schedule of this input. Further, the supply curve of this resource is taken to be an approximation of the value of its social alternatives. Given the usual assumptions for a specified level of employment, a measure of net social benefit can be derived by measuring the area lying between these schedules (*DD* and *SS* in the figure following). Institutional considerations aside, the Paretian level of employment is shown at *OA*, where social benefits are maximized. By imposing a coercive subsidy on this factor, the draft leads the military to "purchase" an excessive amount, *OB*, at the draft supply price $\hat{p}$. The welfare loss associated with this allocation is in excess of the area of the triangle $\alpha\beta\gamma$, since certain other costs do not show up on this diagram. These latter arise from 1) excessive training costs associated with high personnel turnovers; 2) the evasion and avoidance efforts of potential conscriptees attempting to escape the draft; and 3) the cost associated with drafting some individuals whose opportunity costs exceed those of others who are exempted.

Distributional considerations aside, the proponents of voluntarism hold that it would be optimal since under this regime the military is forced to pay the market supply price of this factor instead of the subsidized price, $\hat{p}$. It will be demonstrated that this is not strictly correct since it neglects the possible deadweight burden associated with monopsonistic purchase of volunteers. Such behavior may at first sound unreasonable; but the potential for it is implicit in the assumptions made by voluntarists in their condemnation of the draft together with the evidence concerning the supply function of that factor.

Since the voluntarists hold that the draft leads to an excessive purchase of enlisted men, they implicitly assume that the military treats the budgetary cost of that input under conscription as the actual cost. Does not consistency require that this assumption hold under voluntarism as well?[3] Further, since the empirical evidence indicates that the supply function is upward sloping (see

* The author, assistant professor at the University of Washington, thanks John S. McGee, Yoram Barzel, Lowell R. Bassett, and Allan Hynes for their many helpful suggestions. He also extends his appreciation to the Relm Foundation, whose summer grant for 1969 financed this research as well as other work still in process.

[1] Mark Pauly and Thomas Willet are the most recent and best known of this group. Their paper also surveys past positions on these matters. Other papers in the same volume are useful to read on the subject of efficiency and military procurement.

[2] By defense establishment, I mean not only the military and the Department of Defense, but Congress which acts as an intermediary between the voters and taxpayers and the *DOD*.

[3] Consider, for example, the possibility that the military actually does treat the supply price of enlisted personnel as their cost. It would follow, then, that under voluntarism the optimal quantity would be purchased. It would also be true that under conscription the correct numbers would be drafted. Except for distributional considerations, and perhaps certain other inefficiencies of conscription already discussed, there would be little to choose between the two institutions.

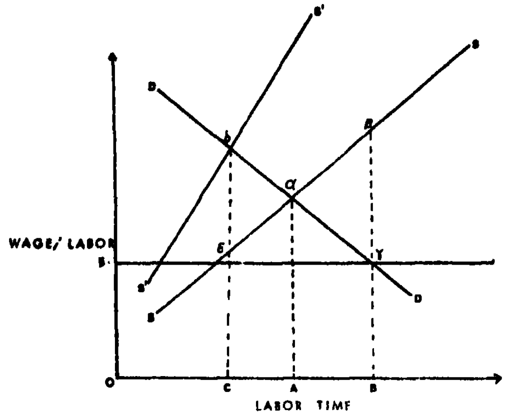


FIGURE 1

Stuart Altman and Alan Fechter, and Walter Oi), monopsonistic purchase is a distinct possibility unless wage discrimination is possible.[4]

This latter phenomenon would bring the marginal budgetary cost in line with the supply price, but it would appear unlikely for two reasons. First is the usual information problem: the determination of individual reservation prices is neither easy nor costless. Second, and more important, is the political constraint: equal wages for identically defined jobs are now the law of the land. Given the unlikelihood of wage discrimination, the relevant incremental factor-cost function to the military would not be the supply schedule, SS, but its marginal, $S'S'$.[5] The resulting monopsonistic allocation, OC, occasions a welfare loss equal in area to the triangle $\alpha\delta\epsilon$.

Although we have some information about the supply function, without some knowledge of the demand schedule and, possibly, some notion of the other costs associated with the draft, it is impossible to specify which inefficiency is more damaging. A rough idea about the magnitude of AB and CA might be helpful. For example, if we could know that AB, the overextension of purchase under conscription, exceeds CA, the underutilization of voluntarism, then a simple assumption that DD and SS are linear (or convex to the horizontal axis) over this range would lend great support to the voluntary military scheme. If, instead, CA were the larger, actual calculation of the welfare triangles $\alpha\beta\gamma$ and $\alpha\delta\epsilon$ would be required, as well as some estimate of the other costs of conscription.

Whether this counter-example points out a serious social cost or is a mere pathological construction based on a doubtful premise would appear an interesting area for further exploration. Such a project would require not only a fair amount of empirical study but, more importantly, a thoughtful examination of our current model of public choice.

[4] In Altman's work the notion of monopsony purchase by a public agency is employed to explain the existence of vacancies. I have recently discovered that Robert Bish and Patrick O'Donoghue have also articulated the problem as does James Buchanan (pp. 89–92).

[5] Empirical evidence suggests that the elasticity of supply is between 1.2 and 1.8 (see Altman and Fechter, and Oi). Thus $S'S'$ must lie from 50 to 85 percent higher than SS.

REFERENCES

S. H. Altman, "The Structure of Nursing Education and Its Impact on Supply," in H. E. Klarman, ed., *Empirical Studies in Health Economics*, Baltimore 1970.

——— and A. E. Fechter, "The Supply of Military Personnel in the Absence of a Draft," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 19–31.

R. L. Bish and P. D. O'Donoghue, "A Neglected Issue in Public Goods Theory: The Monopsony Problem," *J. Polit. Econ.*, forthcoming.

J. M. Buchanan, *Cost and Choice*, Chicago 1969.

W. Oi, "The Economic Cost of the Draft," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 39–62.

M. V. Pauly and T. D. Willett, "Efficiency in Military Manpower Procurement," in J. C. Miller III, ed., *Why the Draft? The Case for the Volunteer Army*, New York 1968.

On the Extension of Input-Output Analysis to Account for Environmental Externalities

By A. O. Converse*

Robert Ayres and Allen Kneese have extended input-output analysis to include the flow of wastes to the environment and the recycle flow of wastes to the production sectors. This was done by adding two sectors, consumption and environment. All "output" from the consumption sector is either recycled or sent to the environmental sector. In addition to this material, the environmental sector also receives flows directly from the various production sectors.

In the Ayres and Kneese model, the flow of waste from the kth production sector to the environment is given by $C_{k0}X_0$ where C_{k0} equals the fraction that comes from the kth sector and X_0 is the total mass of residuals discharged to the environment.[1] Assume that the output of the non-waste commodity from the jth sector increases. This would cause an increase in raw materials required and hence, an increase in X_0. Let Δ be the amount of this increase. The increase in the flow of wastes from the kth sector would be $C_{k0}\Delta$ whether the production in that sector were changed or not. Furthermore, the increase from the jth sector would be $C_{j0}\Delta$ rather than proportional due to the change in the amount of the jth commodity. If such a model were used to distribute a pollution tax, it would be an unfair distribution since the tax would not be directly related to the amount of residuals being discharged to the environment by the particular sector.

In the usual production sectors of the Leontief input-output analysis, the inputs are set by the output from that sector. However, in the case of waste inputs to the environmental sector, the inputs are set by the outputs of *other* sectors. This rather fundamental relationship is absent from the Ayres and Kneese formulation.

I. Waste from the Production Sectors

The model should be revised so that the flow of waste residuals from the kth production sector to the environment is given by $C_{k0}X_k$, where X_k is the flow of the non-waste commodity from the kth sector and C_{k0} is the ratio of the waste residuals to the non-waste commodity in the kth sector. This would overcome the above objections.

In making the above modification one begins to account for the multiproduct nature of production, which is after all, the basic cause of pollution from the production sectors.[2] The above form can account for only one type of waste residual from a given production sector. The model could easily be extended as follows to account for the full range of products.

Let

(1) $\nu_{ij}U_j$ = amount of ith commodity produced or consumed by the jth activity

where:

U_j = extent of the jth activity, as measured by any *one* of its inputs or outputs, or perhaps by the amount of fixed capital invested.[3]

ν_{ij} = amount of the ith product or commodity used or produced per unit

* Associate professor of engineering, Thayer School of Engineering, Dartmouth College.

[1] X_0 is equal numerically to the total amount of raw materials withdrawn from the environment, inventories being neglected.

[2] This aspect is important in other applications of input-output analyses. For example, Alan Manne pointed out the importance of this in his analysis of the petroleum industry.

[3] Perhaps some function of the fixed capital invested would be best as it could account for non-constant returns to scale.

activity in the jth sector, positive for products and negative for inputs.[4]

With this extension one can account for the various types of waste residuals and not merely the total amount. This is obviously very important when trying to evaluate pollution effects. As a matter of fact, the whole treatment effort is based on the assumption that it is desirable to increase the total amount of waste residuals discharged to the environment in order to change the composition.[5] One possible use of an extended input-output simulation would be to test this assumption.

II. Waste from Consumption Activities

The consumption inputs, Y_i, are set either by the planner or the market. During consumption they are not really consumed but merely transformed into wastes.

$$W_k = \sum_i \alpha_{ki} Y_i \tag{2}$$

Whereas the above relationship can account for different types of wastes, Ayres and Kneese lump them all into one item. The increased detail will give rise to greater computational effort but it may well be worth it. For certain applications the detail of the breakdown could be quite modest, e.g., products that are discharged into the atmosphere and those that are discharged into the surface waters. These wastes, W_k, are subject to further processing,[6] hence one must expand equation (1) to

$$X_i = \sum_j \nu_{ij} U_j + \sum_k \gamma_{ik} W_k \tag{3}$$

[4] This corresponds to a stoichiometric coefficient in a chemical reaction. The use of such notations was suggested to the author by Mr. Peter Brooks, University of Queensland.

[5] The total amount is increased because the pollution control treatment requires increased use of resources.

[6] If there really were no further processing, then setting $\gamma_{kk}=1$ and $\gamma_{ik}=0$ for all $i \neq k$, would formally convert all "wastes" into "commodities."

We now distinguish between three types of commodities:

- Goods for consumption, $Y_i = X_i$, assuming that demands are met.
- Materials withdrawn from the environment $= X_i$, when $X_i < 0$.
- Materials discharged to the environment $= X_i - Y_i$, when $X_i - Y_j > 0$.

Conservation of mass requires that

$$\sum X_i - Y_i = 0 \tag{4}$$

all i corresponding to material commodities

It should be noted that the above equality holds only when the flows, X and Y, are expressed in units of mass.

III. Summary

The modification of input-output analysis presented by Ayres and Kneese does not correctly account for the individual waste residues from the various production sectors. A modest change that overcomes this objection is presented. Further modifications that would allow one to account for the various types of waste residues from both production and consumption activities are presented. The need for such detail is caused by the specific activities of the various residues (CO_2 is significantly different from CO). It is noted that pollution treatment while changing the composition of the waste residues does increase the total amount of them. Hence any analysis that considers only the total amount will be unable to evaluate pollution control measures.

REFERENCES

R. U. Ayres and A. V. Kneese, "Production, Consumption, and Externalities," *Amer. Econ. Rev.*, June 1969, *59*, 282–97.

A. S. Manne, "A Linear Programming Model of the U.S. Petroleum Refining Industry," *Econometrica*, Jan. 1958, *26*, 67–106.

Mishan on the Gains from Trade: Comment

By MEL KRAUSS AND DAVID M. WINCH*

In a noted article, Harry G. Johnson developed an analytical apparatus for measuring the welfare gains from international trade in terms of "the goods that could be extracted from the economy in the free trade situation without making the country worse off than it was under protection—some variant of the Hicksian compensating variation (p. 329).[1] In a general equilibrium model, Johnson first divides the welfare effect of a change in commercial policy—the removal of an autarkic tariff on imports, for example—into gross changes in consumers' and producers' surplus, arriving at a *net* welfare effect defined in terms of consumers' surplus; then subdivides this net welfare effect into two separate and distinct components, the first reflecting the *consumption cost* of the tariff; the second, the tariff's cost in terms of deviation from the optimal pattern of production, i.e., the tariff's *production cost*.

This approach to the welfare theory of tariffs has been questioned by E. J. Mishan in a recent article in this *Review* (pp. 1280–82). Though Mishan appears to accept Johnson's breakdown of the welfare effect of trade into consumption and production cost components, he is not willing to go along with Johnson's expression of the gain from trade (or loss from a tariff) as a net gain of consumer's surplus on the grounds that the "... division of the welfare gain from free trade into gains (or losses) of consumer's surpluses offset by losses (or gains) of producer's surpluses is ... arbitrary" (Mishan, p. 1281), and for that reason—and judging from Mishan's argument, for that reason alone—erroneous. While one could hardly disagree that Johnson's expression of the gains from trade as a net gain in consumer's surplus is arbitrary, and indeed must be arbitrary given the impossibility of distinguishing buyer and seller in an exchange model without fiat money, the authors cannot conclude in the same breath that being arbitrary is the same thing as being erroneous.[2] Indeed if one were to extend Mishan's logic beyond the case in point, the whole of the literature that utilizes the concept "numeraire" would have to be discarded on this account alone.[3]

I

The analysis is best begun by seeking a "correct" interpretation of the concept "surplus." Certainly, the terms consumers' surplus and producers' surplus are misnomers; surplus arises from exchange not from production and consumption, and a change in surplus normally derives from a change in the conditions of exchange, not from the conditions of production or consumption. In a general equilibrium context, surplus arises from exchange of factors for products by households or of products for factors by firms. Though, by necessity, there must be at least two parties to an exchange, the surplus arises from the exchange, and not from either side of the bargain. One can isolate consumers' (or buyers') surplus only by pegging the factor side of the household's exchange to the numeraire by assumptions

* Associate professor and professor of economics, respectively, McMaster University, Hamilton, Ontario.

[1] It should be noted that the compensating variation technique is obviously superior to the classical approach of relating the gains from trade to changes in an index expressing one or another concept of the terms of trade.

[2] It should be noted that in this paper the authors address themselves solely to Mishan's Appended Note (pp. 1280–82). Since Mishan himself freely uses the terms "producers' surplus" and "consumers' surplus" in his note (in seeming contradiction to his recomendation in the main text of his article that the term producers' surplus be struck from the economist's vocabulary), the present authors are reluctant to discontinue this practice.

[3] Since the choice of a numeraire in exchange models without fiat money is necessarily arbitrary, and to be arbitrary is to be wrong according to Mishan, one cannot escape the conclusion that all theory developed from models using numeraire is erroneous.

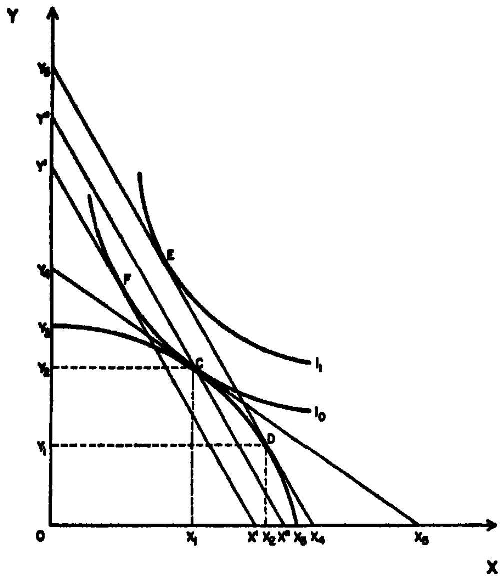


FIGURE 1

about constant income or wage rate. Similarly, producers' (or sellers') surplus is meaningful in isolation only if one pegs all other variables to the numeraire by constant price assumptions. It is quite arbitrary whether a change in the factor to product price ratio is held to yield the household a consumer's surplus (or buyer's rent) as buyer of the product, or a producer's sur-

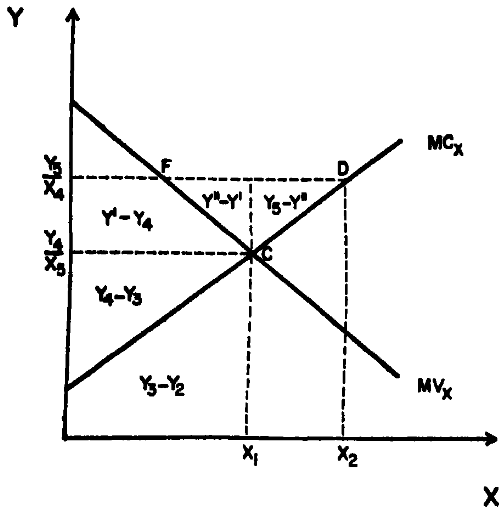


FIGURE 2

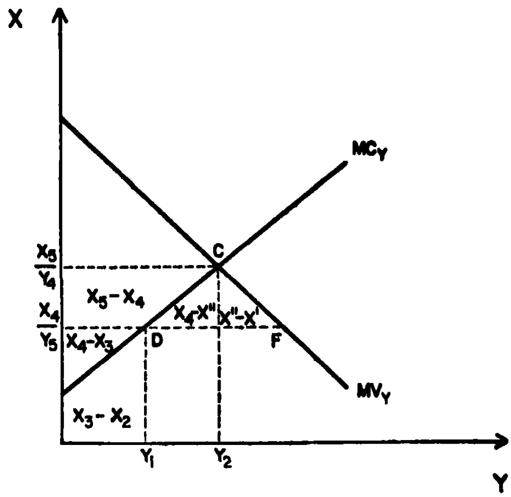


FIGURE 3

plus (or seller's rent) as a seller of a factor. Similarly, it can yield the firm a surplus (profit) as buyer of factor or seller of product. Since in a competitive equilibrium, there is no pure profit (firms' surplus), all surpluses can be attributed to either factor sellers or product buyers depending on the choice of numeraire.

II

The transformation curve x_5y_5 of Figure 1, derived from the familiar Edgeworth-Bowley efficiency box (not shown), is combined with a set of comparable community indifference curves I_0I_1, to yield in Figure 2, a supply (marginal opportunity cost) curve of the export commodity x and a compensated demand curve for x (marginal valuation curve), y being the numeraire; in Figure 3, a supply curve of the import commodity y with a compensated demand curve for y, x being the numeraire. According to the numeraire chosen (an arbitrary choice and one that makes no real difference to the result), the movement from self-sufficiency to free trade (from point C to D) can be shown alternatively as a movement from the price ratio Y_4/X_5 to Y_5/X_4 in Figure 2, or X_5/Y_4 to X_4/Y_5 in Figure 3 (i.e., in terms of numeraire x or numeraire y). In Figure 2 where y is the numeraire, producer's surplus rises and consumers' surplus falls. This is because the total payment to factors expressed in y rises

while the purchasing power of y in buying x falls (the price of x in terms of y rises). Similarly, the movement from autarky to free trade can be shown in Figure 3 where x is the numeraire. Here the total payment to factors expressed in x falls while the purchasing power of x in buying y rises (the price of y in terms of x falls). The sum of the two changes (Y_5-Y') or (X_4-X') is the total welfare effect of moving from I_0 at point F to I_1 at point E in Figure 1. Whether this is represented as a net gain in producers' surplus (Figure 2) using y as the numeraire, or a net gain in consumers' surplus (Figure 3), using x as the numeraire, is as arbitrary as the choice of the numeraire itself. But being arbitrary doesn't make it erroneous. To paraphrase Marshall, it is a matter of indifference which numeraire is used to express a given surplus.

There can be no doubt that Johnson was aware of the numeraire problem inherent in general equilibrium models without fiat money, and thus to the potential misuse of his technique by others in double counting the welfare gain from trade as excess consumers' surplus on x *plus* excess producers' surplus on y.[4] The two measures are neither separable in the sense that they have real (i.e., utility or welfare) significance in the absence of a numeraire, nor are they addable in the sense that their sum accurately portrays the gain that accrues to the community from the opportunity to trade. Being alternatives, either is sufficient for the purpose of measuring the change in welfare. They are alternative depictions of the same thing, not separate components of a single whole.

A final note relates to Mishan's suggestion that the welfare gain from trade is *best* measured as a single compensating variation at the new international price ratio—as Y_5Y' or X_4X'—and interpreted as an exact measure of the gains for the community as a whole in moving from the consumption possibilities presented by X_3Y_3 to the new consumption possibilities presented by X_4Y_5, with no mention being made of separating this measure into consumption cost and production cost components. While Mishan's overall measure is correct,[5] indeed it is identical with Johnson's, the latter's technique of separation has the substantial advantage of highlighting the sources of the welfare gain, i.e., the gain from substituting lower cost for higher cost goods in consumption, and the gain from diverting resources from direct higher cost to indirect lower cost production of goods that can be imported from the world market. The failure to make use of this empirically relevant and theoretically illuminating analytical distinction is simply to throw out useful analysis for the sake of misplaced methodological purism.

REFERENCES

J. Bhagwati and H. G. Johnson, "Notes on Some Controversies in the Theory of International Trade," *Econ. J.*, Mar. 1960, *70.*, 74–93.

H. G. Johnson, "The Cost of Protection and the Scientific Tariff," *J. Polit. Econ.*, Aug. 1960, *68*, 327–45.

E. J. Mishan, "What is Producer's Surplus?," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1269–82.

[4] This is evident from Johnson's discussion of the changes in the terms of trade that might be induced by moving from a restricted trade to free trade position, and the effect this would have on the measure of the gains from freer trade, i.e., the quantity of goods that could be extracted from the economy, leaving it as well off under free trade as it was with the tariff. On p. 330, Johnson clearly states that the terms of trade gain or loss will differ according to whether the compensating variation is effectuated with exportables or importables. If exportables are used, i.e., if x is the numeraire, the extraction results in a terms of trade gain for the extracting country; if, on the other hand, importables are used, the result is a terms of trade loss. Cognizance of this problem clearly implies cognizance of the general problem of numeraire in general equilibrium models without fiat money, and thus the dangers of double counting. Also, see Jagdish Bhagwati and Johnson, where three alternative measures of the gains from trade are developed, each using a different numeraire.

[5] Mishan's triangle *EHG* of his Figure b (p. 1280) is equivalent to our triangle *DCF* in Figure 3, and the sum of Johnson's two triangles *GHJ* and *DEF* in his Figure 2 (p. 331). The difference between Mishan and the present authors (and Johnson) is that Mishan contends that the identification of triangle *EHG* as a net gain of consumers' surplus over a loss of producers' surplus is erroneous. It should be noted that Mishan's position on this issue appears is no way related to his general condemnation of the concept "producers' surplus" expressed in the main text of his article.

Mishan on the Gains from Trade: Reply

By E. J. Mishan*

Apart from some conventional sniping from behind improvised footnotes, the counterattack mounted by David Winch and Mel Krauss has the apparent aim of recapturing for Harry Johnson the right to continue to use producers' surplus in his analysis of the welfare effects of tariff protection. Since they do not appear to dispute the main thesis developed in the text of my paper (that producers' surplus is either a misnomer or a conceptual error), their gallantry in attempting to secure a special dispensation for the use of producers' surplus in the gains-from-trade case also involves a certain awkwardness; one that is, not surprisingly, reflected in their arguments. Broadly speaking, their tactics are first to churn up a terminological smokescreen so as to obscure traditional notions of consumers' surplus and rents, and then, while the innocent reader is blinking at the resulting swirl, to produce for him concepts guaranteed to be slippery enough to meet any contingency.

I

By adapting Johnson's tariff argument to a special autarky-free trade situation, I affirmed that under special conditions—zero income elasticity of importables, Pareto-comparability of community indifference curves, and the adoption of the free trade terms of trade to measure welfare gains—there is a clear mathematical equivalence of the welfare gain of removing a tariff when measured in terms of a vertical (or horizontal) distance on the transformation-indifference curve construction and when measured in terms of an area bounded by the derived marginal curves. The identification by Winch and Krauss of these ways of measuring the net gain of moving from autarky to free trade does not, therefore, add anything to the argument in my Appended Note. Contrary to what they seem to allege, I did not repudiate Johnson's breakdown of the welfare effects because they were arbitrary with respect to *measure* (being measurable, in the two-good case, in terms either of x or y) but simply because they were arbitrary with respect to *concepts*; by which I refer to Johnson's interpretation of the net community gain as an excess of the gain of consumers' surplus over the loss of producers' surplus or, equally invalid, as an excess of the gain of producers' surplus (in terms of the other good) over the loss of consumers' surplus.[1]

The issue between Winch and Krauss and myself is, therefore, simply this: that they believe the *net* gain from trade can be properly interpreted as an excess of a gain of consumers' over a loss of producers' surplus, or the reverse, and I do not.

In their Section I on the subject of surpluses, "consumers' surplus" and "producers' surplus" are asserted to be misnomers, on the grounds that "surplus arises from exchange, not from the conditions of production or consumption." Consistency of interpretation, however, requires that a measure of surplus be regarded as a measure of the *change* in a person's welfare irrespective of how it is brought about—whether an effective constraint is removed, whether a change in one or more prices favor him, whether the conditions of his work or his environment improve or deteriorate, or whether his wife leaves him. As has already been pointed out elsewhere (J. R. Hicks), the traditional division into a consumer's surplus (in which, narrowly conceived, a per-

* London School of Economics.

[1] In their fn. 2, the authors write: "Since Mishan himself freely uses the terms producers' surplus and consumers' surplus in his note (in seeming contradiction to his recommendation in the main text . . .) the present authors are reluctant to discontinue this practice." As a *non sequitur* this is also ingenuous. It should have been obvious that, in paraphrasing and extending Johnson's agrument, prior to registering my dissent, there was some provisional convenience in employing his own terminology.

son's welfare is affected by changes in one or more product prices, all other product and factor prices constant) and rent (in which, narrowly conceived, a person's welfare is affected by changes in one or more factor prices, other factor prices and all product prices remaining constant) has occasional convenience in welfare analysis—although it should be borne in mind, always, that any gain or loss resulting from a complex alteration of both factor and product prices can, in principle, be measured in terms of income, or in terms of one or more of the constant-priced goods.

Their further pronouncements on this subject add nothing to this, while subtracting a great deal in the way of clarity. Thus producers' surplus is smuggled in again—in a rather casual way, so as not to give offence. For they talk of a 'producer's surplus' (or seller's rent) as accruing to the seller of a factor. From which one might reasonably infer that they would like a licence to use producer's surplus while referring to the rent of a factor. Reluctantly one sympathizes with their cunning, since a vindication of Johnson's use of producers' surplus looks to be an impossible task without injecting occasional doses of ambiguity into the argument. We are, therefore, further informed that a change in the factor-product price ratio can also yield the firm a surplus, which is profit, though in competitive equilibrium there is no firms' surplus, since any surplus must be attributed to factors or consumers. But which it is to be depends, according to the authors, on the choice of numeraire.

Given our traditional definitions, however, it depends on what occurs. If a firm disposes of its profits by lowering the price of its products, *ceteris paribus*, it generates a consumers' surplus. If, on the other hand, it does so by raising the price of its factors, it generates an increase in rents. If, however, we are dealing with the economy as a whole, and the only information we have, in the two-good model, is that the price of one good has risen and that of the other has fallen (or one risen in terms of the other), then we cannot, in general, say either that consumers' surplus has increased or decreased, or that rents have increased or decreased. (We can say nothing about a producers' surplus which—unless it is to be a synonym for Knightian profit, in which case it is zero—does not exist.) For all that, the authors are determined to deliver to us a producers' and consumers' surplus cut out of the gains from trade. Let us see what surprises they come up with.

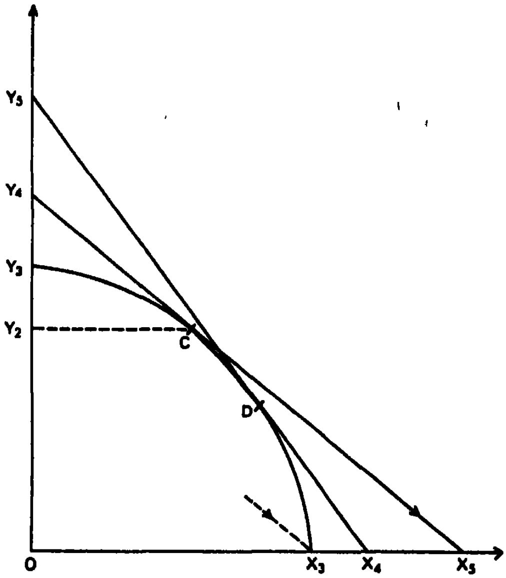


Figure 1′

II

Now it transpires that the crux of their argument, for retaining the use both of a producers' and a consumers' surplus in measuring the welfare gain from trade, turns only on relative price-changes, and on *product* prices at that. It can be summarized with reference to my Figure 1′ (which includes only the features of their Figure 1 that are necessary to this part of their argument).[2]

[2] Their Figures 2 and 3 are used to depict their conclusions, which conclusions depend only on the fact of a change in relative product prices. Nor is the movement from I_0 to I_1 (measured in terms of Y_5X_4) a part of their demonstration that there is both a producers' surplus and a consumers' surplus in terms of one good or the other. This measurement of gain from I_0 to I_1 is, as I argued, a net gain of community welfare and can, as I asserted, be measured in terms of either x or y.

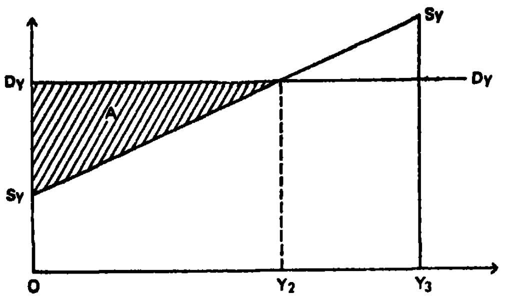


FIGURE 3′

Consider, then, the movement from C to D, along with the accompanying change in the product terms of trade—these being indicated by the shift in the tangential straight lines Y_4-X_5 to Y_5-X_4; and let us forget about producers' surplus inasmuch as the authors' argument, in the first paragraph of their Section II, becomes strictly in terms of factor rents. Have rents changed? Has consumers' surplus changed? Yes, they answer. For a) if we take product y as the numeraire, total rents (i.e., the rents of the two factors A and B) rise from Y_4 to Y_5 whereas total consumers' surplus falls because the price of x rises in terms of the numeraire y. On the other hand, according to the authors, b) if we take product x as the numeraire, total rents fall from X_5 to X_4, whereas consumers' surplus rises because the price of y has risen in terms of the numeraire x. This apparently is the logic that produces an excess of producers' surplus over the loss of consumers' surplus for case a), and the reverse, an excess of gain of consumers' surplus over the loss of "producers' surplus" for case b). So it all depends, after all, on the numeraire.

The reader will observe in passing: i) that these relative changes occur simply in moving from C to D, regardless of whether such movement is from autarky to trade, or from trade to autarky, or the result of a change in tastes or of a change in the distribution of the product; ii) that, at the relevant terms of trade, batch C is valued at Y_4 or X_5 while batch D is valued at Y_5 or X_4. Therefore, when the movement from C to D is accompanied by a rise in the price of x relative to that of y, the community *as a whole*—whether they are regarded as consumers or producers[3]—looks better off in terms of y and looks worse off in terms of x. From this change, it is not possible to infer an opposition of welfare as between consumers as a whole and earners as a whole; and iii) that if x is A-intensive then, as the Stolper-Samuelson theorem affirms (for two-input production functions homogeneous of degree one), the owners of factor A are *absolutely* better off in the movement from C to D, whereas the owners of factor B are *absolutely* worse off. But no means are devised in these constructs for determining whether, *on balance*, total real rents have increased or not. Consequently it is just not possible to establish that the real increase of welfare which (given Pareto-comparability of I_0 and I_1) occurs in the movement from C, the autarkic position, to D, the free trade position, is the resultant of a real increase in rents that is greater than a real decline in consumers' surplus, or the reverse.

Indeed, as I pointed out, the welfare gain is represented on the original diagram as a measure only (at the international terms of trade) of moving from community indifference curve I_0 to community indifference curve I_1. Neither the welfare of factors or of firms can be distinguished or isolated from the original indifference-transformation construction, since they are not there to begin with.

What then are we to make of the plausible-looking areas in Winch and Krauss's Figures 2 and 3?

III

A useful clue is provided if we begin with a Robinson Crusoe economy, in which point C is chosen on the transformation curve[4] of my Figure 1′, with the tangential line Y_4-X_5

[3] In the sense that if the terms of trade consistent with point D could be maintained for any amount of x exchanged for y, the community could end up with OY_5 of y, whereas with the terms of trade consistent with point C the community could end up with only OY_4 of y.

[4] No reason need be given for its convexity, nor for how it comes into being. To Crusoe, the locus of alternative combinations of x and y is to be accepted as a datum.

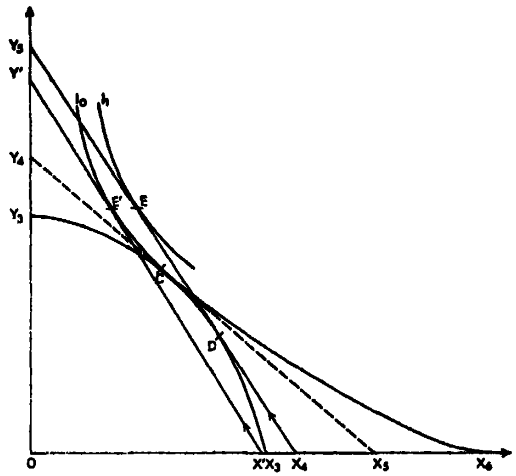


FIGURE 1''

representing Crusoe's indifference curve. Taking the first derivative of the transformation function with respect to y, we construct Figure 3' (comparable with Figure 3 drawn by Winch and Krauss).

How do we interpret the shaded triangle A? The schedule S_yS_y, the derivative of the transformation curve with respect to y, begins from a position in which Crusoe holds only x and shows the increasing amounts of x that have to be surrendered for each additional unit of y. The height of the horizontal line D_yD_y represents the unchanged marginal value to Crusoe (in terms of x) of these additional increments of y. The area of the triangle A, therefore, measures the welfare gain of Crusoe's taking OY_2 of y on these terms.[5] If we have to put a label to the welfare gain, we should call it a buyer's or a consumer's surplus in deference to the traditional commonsense definition that harks back to Marshall—being the difference between what he (Crusoe) is willing to give up for OY_2 of y, and what he has to give up for it. And this

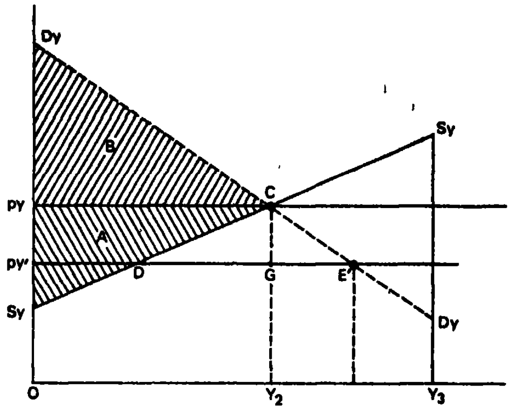


FIGURE 3''

area which is more suggestively regarded as a consumer's surplus—and *not* a producers' surplus as Winch and Krauss would have it—corresponds in my Figure 1' to the horizontal distance X_5-X_3. This is the maximum amount of x that Crusoe would pay in order to have the terms of trade Y_4X_5 since they enable him to move from X_3 to C.[6]

The reader will readily appreciate that no difference is made to this analysis if the indifference curve touching C took on the familiar concave shape as shown by I_0 in my Figure 1'' and, in consequence, the D_yD_y curve (the marginal valuation of y in terms of x) were downward sloping as shown in Figure 3''. Crusoe's gain in welfare, still regarded as a consumer's surplus, would then be the area of the two shaded triangles A and B, a total area that corresponds to the distance X_6-X_3 in Figure 1''.

Let Crusoe now be faced with world terms of trade Y_5X_4 in Figure 1''. He does the best for himself by producing the combination D and exporting as much of x as is necessary to enable him to consume combination E on the I_1 curve. The gain in moving from the consumption of C on the I_0 curve to the consumption of E on the I_1 curve can be measured either in terms of x or of y, and,

[5] If Crusoe begins with OX_3 of x and can transform x into y at a constant rate equal to his constant marginal valuation, he gains nothing from exchanging y for x. It is just because he need give up less of x than this for the initial units of y that enables him to increase his welfare, an increase that can be measured as the horizontal (or vertical) distance between the straight line indifference curve passing through X_3 and the parallel indifference curve passing through X_5.

[6] This welfare gain can also be measured, at the same terms of trade, in terms of y. Alternatively, we could measure a different welfare gain (in terms of x or of y) of being presented with these same terms of trade if, instead, we suppose Crusoe to begin at Y_3, having only y and no x.

moreover, at either the old terms of trade or at the new. If the gain from trade is measured at the new international terms of trade, and in terms of x, the movement from I_0 to I_1 can be measured as a compensating variation X_4-X'; i.e., as the maximum amount of X that Crusoe is willing to sacrifice in order to have the privilege of trading at the new international terms of trade.

If we let the "income" elasticity of demand for y be zero, so that E' on I_0 is on a line through E that is parallel with the x-axis of Figure 1″, then D_yD_y in Figure 3″ is the compensated demand curve. The compensating-variation measure of gain from a lower price of y in terms of x, p_y', is accurately depicted as equal to the triangle DCE'.

The interpretation of this gain from trade is also straightforward, and has no reference to a producers' surplus, or to rents of factors. Crusoe will continue to acquire y from domestic sources until point D is reached on his S_yS_y curve. Beyond that point, it is cheaper for him to buy y abroad at a fixed price p_y' (the amount of x he gives up being readily accepted as exports). And he continues to buy y at this international price p_y' until he reaches point E', beyond which his marginal valuation of y falls below this p_y' price.

This area of gain, DCE', can be split into two triangles: the area of the first triangle DCG being that part of the gain from being able to buy some part, DG, of Crusoe's original consumption of y at an international price, p_y', that is lower than the domestic price, p_y. The second triangle GCE' is the additional gain arising from the extra quantity, GE' of y, that Crusoe will consume as a result of the lower international price p_y'.

The identification by Johnson of the gain represented by the first triangle, DCG in Figure 3″, as "the increase in the value of production," and that of the second triangle, GCE' in Figure 3″, as "the reduction in the cost of consumption," does, perhaps, make some sort of sense, for what the distinction is worth. For the area of the first triangle depends on the shape of the marginal transformation curve S_yS_y and can, therefore, be associated with the conditions of production, while the area of the second triangle depends upon the shape of the compensated demand curve D_yD_y and can, therefore, be associated with the conditions of consumption.

What *cannot*, however, be inferred from the derived triangle DCE' in Figure 3″—whether we remain with Robinson Crusoe or whether, instead, we have Figure 1″ refer to the community—is that it is the resultant of a gain of consumers' surplus over a loss of producers' surplus. For, as we have seen—regardless of whether we obstinately think of producers' surplus in terms of profits or of rents to factors—there is no real loss of producers' surplus in moving from the production of C to that of D. There is only the compensating variation measure of community gain in moving from I_0 to I_1,[7] a gain that is represented by the triangle DCE' in Figure 3″.

IV

Finally, a word on the alleged empirical usefulness of Johnson's technique.[8] Although the analysis did not lack in sophistication, a cursory reading at least gives the impression that its value is more heuristic than empirical. Allowing that a measurement of gains (or losses) of tariff-reductions proceeds by a consideration of importables only, calculation of the net gain in the two-good case would require a supply curve and a compensated demand curve of x in terms of y.

[7] Their "excess producers' surplus"—triangles $Y''-Y'$ and Y_3-Y'' in their Figure 2—is also equal to a compensating variation, in term of y, of moving from the consumption of C on I_0 and that of E on I_1, this area corresponding to the vertical distance Y_3-Y' in their Figure 1.

[8] It should be noted, however, that their alleged "substantial advantage of highlighting the sources of welfare gains, i.e., the gain from substituting lower cost for higher cost goods in consumption, and the gain from diverting resources from direct higher cost to indirect lower cost production of goods that can be imported from the world market." which can be translated into Johnson's division of the net gain into a "reduction in the cost of consumption" and an "increase in the value of production," (that correspond to triangles DCG and GCE' in my Figure 3″) is not something I take issue with. The issue, to repeat, is the interpretation of the net gain, measured by the combined area of these two triangles, as one resulting from an excess in the gain of consumers' surplus over the loss of a producers' surplus.

In the real world estimates of demand and supply curves are in terms of money prices, with money prices of other goods constant. But, first of all, once the analysis is extended to a number of goods, it is no longer reasonable to assume upward-sloping supply curves even if the model is restricted to two factors. Assuming production functions are homogeneous of degree one (or less), at least, it is no longer possible to infer unambiguously that an increase in the demand for a good will raise its long-run equilibrium supply curve either absolutely or relatively.[9]

Again the estimates of the kind of demand functions that are required to calculate the gains or losses arising from a reduction of the tariffs on a number of importables are not easy. If x_1, x_2, . . . x_n, are importables that are, in some degree, substitutes, the gain (or loss) from a simultaneous removal of their tariffs is not to be measured by adding together the apparently relevant areas under the more usual *ceteris paribus* demand curves for each of the x_1, x_2, . . . , x_n importables. Following J. R. Hicks' analysis of 1956, the total gain is to be calculated by taking the tariff-removals in sequence; adding the gain from removing the x_1 tariff alone, the tariffs on the remaining $(n-1)$ goods unchanged, to the gain from removing the x_2 tariff alone when the x_1 tariff is already removed but the tariffs remaining on the other $(n-2)$ goods, and so on until the x_n tariff.

Demand curves estimated for these conditions are not, however, readily available. And although I am inclined to agree with Johnson that, for Western countries at least, the gain from removing tariffs is not likely to be very significant in terms of *GNP*, such conclusions depend much more on rough guesses than on any careful application of the apparatus he skillfully designed in his 1960 paper.

[9] If there are n-goods, x_1, x_2, . . . , x_n, ranked thus in order of increasing A-intensity, a shift of demand, from say, x_4 to x_5, raises the price of factor A (on the common assumption of some inelasticity in its supply) relative to B (though raising it by less the larger the number of A-using goods having some factor substitutability). This rise in the relative price of A raises the supply price of x_n relative to that of x_{n-1} . . . relative to that of x_5 relative to that of x_4 . . . relative to that of x_1—assuming, always, the A-intensity ranking remains unaltered. Given a constant stock and velocity of money, we can say nothing of the money price of x_5, however, without further specification. We can deduce for certain only that the money price of x_n will rise. On the other hand, a shift of demand from good x_8 to x_5 lowers the price of A relative to B, and lowers the price of x_n relative to that of x_{n-1} . . . relative to that of x_8 . . . relative to that of x_5 . . . relative to that of x_1. Again, without further specification, we can say nothing of the change in the money supply price of x_5. We can deduce only that the money supply price of x_n will fall.

REFERENCES

J. R. Hicks, A *Revision of Demand Theory,* Oxford 1956.

H. G. Johnson, "The Cost of Protection and the Scientific Tariff," *J. Polit. Econ.*, Aug. 1960, *68*, 327–45.

E. J. Mishan, "Rent as a Measure of Welfare Change," *Amer. Econ. Rev.*, June 1959, *49*, 386–95.

Profit Constrained Revenue Maximization: Note

By RICHARD ROSENBERG*

In recent issues of this *Review*, the profit constrained, revenue maximizing oligopoly model introduced by William J. Baumol has been both expanded and applied. (See articles by Robert Haveman and Gilbert DeBartolo (1968, 1970), C. J. Hawkins, Milton Z. Kafoglis, Kafoglis, and Robert C. Bushnell, and David J. Smyth.) The attention paid to this model can be interpreted as a symptom of the general feeling of dissatisfaction with regard to the lack of an acceptable theoretical explanation for much of oligopoly behavior. While Baumol has made an important advance by introducing a two-dimensional objective function for oligopoly firms, he has, unfortunately, also specified a particular form for the objective function which turns out to be untenable.

Unlike the use of constraints in defining an environment or in determining the subset of the choice set which is attainable, the use of constraints in an objective function has clear implications for both the method of arriving at the decision maker's ordering and for his behavior. To postulate that firms seek to maximize sales revenue, subject to a profit constraint, implies that firms order various outcomes (each outcome is a combination of a certain level of profits and a certain level of sales revenue) in a lexicographic manner as illustrated in Figure 1. Each of the lines in the figure is a behavior line rather than an indifference curve. This implies that the firm orders possible outcomes in the following way: 1) for any two outcomes, both of which have profits below the constraint, the outcome with the larger profit is preferred (B is preferred to A) regardless of the associated levels of sales revenue. If both have the same level of profit, the outcome with the larger sales revenue is preferred (C is preferred to B); 2) for any two outcomes, both of which have profits equal to or greater than the constraint, the outcome with the greater sales revenue is preferred, (E is preferred to D). If both have the same level of sales revenue, the outcome with the larger profit is preferred (D is preferred to F).

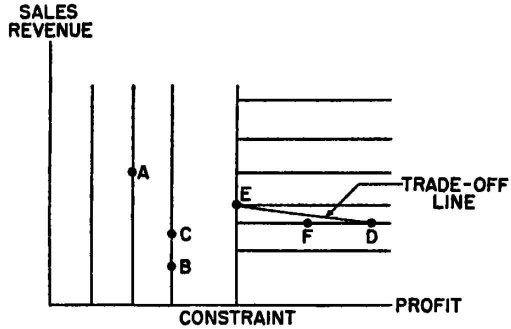


FIGURE 1

Moreover, since the model assumes that dollars of profit can always be converted into additional sales revenue through promotional activities, if the attainable set of outcomes includes one outcome for which profits exceed the constraint, then the attainable set will also include an outcome at which profits are equal to the constraint and sales revenue will be increased. In the figure, if the firm can attain a position such as D by profit maximizing behavior, it can and will move along a trade off line such as DE until it reaches the profit constraint at E. The firm makes this trade no matter how small the increase in revenue relative to the decrease in profit.

In conventional terminology from the theory of demand, the firm is thus assumed to have a Marginal Rate of Substitution of sales revenue for profit (the amount of profit it is willing to give up in order to receive an extra unit of sales revenue) which is infinite so long as profit exceeds the constraint, and which is always equal to zero so long as profit is below the constraint.[1] While

* Assistant professor of economics, Pennsylvania State University.

[1] This implication was recognized as long ago as 1965 by Armen A. Alchian.

such an ordering is conceptually possible, there does not seem to be any economic rationale for supposing that such a strong and unstable preference pattern should exist. While it may prove fruitful to postulate that firms are not motivated solely by the desire for greater profits, the constrained maximization approach is not a satisfactory method for embodying this notion.

REFERENCES

A. A. Alchian, "The Basis of Some Recent Advances in the Theory of the Management of the Firm," *J. Ind. Econ.*, Nov. 1965, *14*, 30–41.

W. J. Baumol, *Business Behavior, Value and Growth*, rev. ed., New York 1967.

R. Haveman and G. DeBartolo, "The Revenue Maximization Oligopoly Model: Comment," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1355–58.

———, "The Revenue Maximization Oligopoly Model: Reply," *Amer. Econ. Rev.*, June 1970, *60*, 433–34.

C. J. Hawkins, "The Revenue Maximization Oligopoly Model: Comment," *Amer. Econ. Rev.*, June 1970, *60*, 429–32.

M. Z. Kafoglis, "Output of the Restrained Firm," *Amer. Econ. Rev.*, Sept. 1969, *59*, 583–89.

——— and R. C. Bushnell, "The Revenue Maximization Oligopoly Model: Comment," *Amer. Econ. Rev.*, June 1970, *60*, 427–28.

D. J. Smyth, "Sales Maximization and Managerial Effort: Note," *Amer. Econ. Rev.*, Sept 1969, *59*, 633–34.

Behavior of the Firm Under Regulatory Constraint: Note

By ISRAEL PRESSMAN AND ARTHUR CAROL*

A recent article by Akira Takayama discusses an earlier paper by Harvey Averch and Leland L. Johnson on fair rate of return regulation of public utilities. Although Takayama (p. 255) agrees with Averch and Johnson's general conclusions "that a firm will tend to increase its investment with the introduction of an active constraint" on its rate of return, he criticizes the A-J argument as being "confusing, ambiguous, and in error." Takayama then attempts a clarification, and presents a new formulation which leads to the A-J result quoted above.

This comment will discuss several of Takayama's criticisms in addition to showing that the so-called "A-J Effect" cannot be derived from the basic assumptions made by both Averch and Johnson and Takayama.[1] We will show that the very assumptions used to prove the A-J Effect, by defining the region of λ, require an assumption that the A-J Effect exists in the first place.

I. The Model

Consider a monopoly employing two inputs, capital (x_1) and labor (x_2) to produce a single homogeneous output (z). The firm faces a production function

$$z = z(x_1, x_2),$$

defined on $x_1 \geqq 0$, $x_2 \geqq 0$, having positive first-order derivatives and satisfying $z(0, x_2) = z(x_1, 0) = 0$. The price ($p$) is related to the output by the inverse of the demand function, i.e.,

$$p = p(z) \tag{1}$$

The profit is defined by

$$\pi = pz - r_1x_1 - r_2x_2 \tag{2}$$

with r_1 and r_2 the factor prices presumed constant. The revenue, pz, is assumed to be concave. The regulatory constraint on the rate of return is given as[2]

$$\frac{pz - r_2x_2}{x_1} \leq s_1 \tag{3}$$

The problem then is to maximize (2) subject to (3). The Lagrangian $L(x_1, x_2, \lambda)$ is formed and the Kuhn-Tucker necessary conditions for a maximum at $\bar{x}_1$, $\bar{x}_2$, $\bar{\lambda}$ are given.[3]

II. Analytical Aspects

Comments on Takayama's Criticisms

Consider first the question of whether λ, the Lagrange multiplier, has a range $0 \leq \lambda < 1$. Averch and Johnson (p. 1055) prove that $0 < \lambda < 1$ by noting three things: first, "that $\lambda = 1$ if, and only if, $r_1 = s_1$;" second, since $s_1 > r_1$, $\lambda \neq 1$; and finally, λ varies continuously. Takayama criticizes this proof as to the question of continuity of λ with s_1. He then indicates conditions under which λ will be continuous.[4]

To derive the conditions of continuity, Takayama writes an equation for λ as

$$\lambda = \frac{G_K - r}{G_K - s} \tag{4}$$

* The authors are, respectively, assistant professor of operations research at the Polytechnic Institute of Brooklyn and associate professor of economics at the University of Hawaii. Pressman is also a member of the Management Sciences Division of the American Telephone and Telegraph Company. The authors are indebted to the referees and all those who made helpful comments on earlier drafts of this paper. The views expressed here are solely our own.

[1] Both Averch and Johnson and Takayama define these assumptions as a concave revenue function and that s_1, the maximum allowed rate of return, is strictly greater than r_1, the cost of capital.

[2] See Averch and Johnson, pp. 1054–55, for discussion of the assumptions leading to this formulation of the constraint.

[3] Averch and Johnson, p 1055.

[4] We note that Averch and Johnson fail to prove the "if" part of their statement, i.e., $r_1 = s_1$ implies $\lambda = 1$. In fact, no analytic proof known to the authors has been given in the literature to support this assumption. An alternate proof of the range of the Lagrange multiplier is presented by William Baumol and Alvin Klevorick. Their proof, however, depends on the same basic assumptions required by Averch and Johnson, i.e., a concave revenue function and $s_1 > r_1$.

assuming $G_K - s \neq 0$,[5] and noting that the r in equation (4) is the r_1 in the A-J notation.

We note that since $s > r$, the value of λ will depend on whether G_K is greater than both s and r, less than both s and r, or greater than r but less than s. If G_K is greater than both s and r, then $\lambda > 1$. If G_K is less than both s and r, then $0 < \lambda < 1$. These results are entirely possible if the constraint is an equality and Lagrange Multiplier optimization techniques are employed. Thus equation (4) above gives no clue as to the value of λ.[6]

Finally, Takayama criticizes Averch and Johnson for assuming that the marginal-revenue-product-of-capital curve (*MRPK*) does not shift. He claims (p. 257) that this assumption "is not true in general," and "involves an error of confusing the movement along the curve with the shift of the curve." If, however, this criticism is valid and the *MRPK* does actually shift with changes in labor and capital, then the assumption "G_K is a continuous function of L^* and K^*" (p. 259), (which is made to define the conditions under which λ^* is continuous) is not necessarily valid. In addition we know of no evidence mathematically or empirically that "L^* and K^* are continuous functions of s."

Comments on the A-J Effect

A major objection to both Averch and Johnson and Takayama is the assumption that so-called A-J Effect can be derived from the basic assumptions. Actually, the two assumptions[7] needed to prove the A-J Effect (by defining the region of λ) require the existence of this effect. Consider the A-J formulation of the problem given above, i.e.,

$$\text{(5)}\qquad \begin{aligned} &\max \pi = pz - r_1x_1 - r_2x_2 \\ &\text{s.t. } pz - s_1x_1 - r_2x_2 \leq 0 \\ &x_1, x_2 \geq 0 \end{aligned}$$

The essential discussion of A-J is when the constraint is operative,[8] i.e.,

$$\text{(6)}\qquad pz - s_1x_1 - r_2x_2 = 0$$

Thus, a solution to the problem (by inspection) is to increase x_1 (capital) to a value greater than the unconstrained value x_1^0, say to x_1^*. This would assure a reduction of the unconstrained rate of return s_0 to the regulated value s_1. That this is the solution to problem (5) has also been shown geometrically by Eugene Zajac. Therefore, we wish to investigate only the case $x_1^* > x_1^0$. Writing the Lagrange function for problem (5) we have

$$\text{(7)}\qquad \begin{aligned} L = (1 - \lambda)pz + (\lambda s_1 - r_1)x_1 \\ - (1 - \lambda)x_2r_2 \end{aligned}$$

Taking partials with respect to x_1, x_2, and λ we get

$$\text{(8)}\qquad \alpha G^*_{x_1} + \lambda s_1 - r_1 = 0$$

$$\text{(9)}\qquad \alpha G^*_{x_2} - \alpha r_2 = 0$$

$$\text{(10)}\qquad pz - s_1x_1^* - r_2x_2^* = 0$$

where $\alpha = 1 - \lambda$. Since we require that $r_1 < s_1$ we have

$$\text{(11)}\qquad r_1 = \alpha G^*_{x_1} + \lambda s_1 < s_1$$

or

$$\text{(12)}\qquad \alpha G^*_{x_1} < (1 - \lambda)s_1 = \alpha s_1$$

Thus

$$\text{(13)}\qquad G^*_{x_1} < s_1 \text{ implies } \alpha > 0$$

Although equation (13) allows for $G^*_{x_1} < 0$, the possibility of having both $G^*_{x_1} < 0$ and $\alpha < 0$ does not occur since $\alpha < 0$ implies $\lambda > 1$ which then implies, from equation (11), that $r_1 > s_1$, a nonallowable condition. If we assume that $G^*_{x_1} < 0$, then $r_1 > G^*_{x_1}$ by definition. Then, since $r_2 = G^*_{x_2}$, either $\left(p + z\dfrac{dp}{dz}\right)$

[5] We note that G_K is equivalent to Averch and Johnson's

$$\left(p + z\frac{dp}{dz}\right)\frac{\partial z}{\partial x_1}$$

In addition, Takayama uses y, K, L, r, w, and s where Averch and Johnson use z, x_1, x_2, r_1, r_2, and s_1.

[6] Takayama's further contention that $G_K - s = 0$ implies $s = r$ is correct; however, equation (4) in no way includes the assumptions $s > r$, nor does it define a value of λ when $G_K - s = 0$.

[7] a) concave revenue function, and b) $s_1 > r_1$.

[8] Takayama, p. 256, fn. 8.

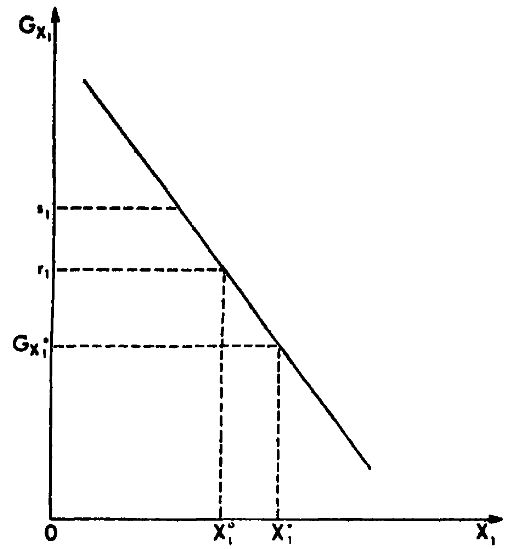


FIGURE 1

and $\partial z/\partial x_2$ are negative, *or* $\partial z/\partial x_1 < 0$. Thus the optimum operating point occurs where one of the marginal products is negative, i.e., in the region of economic inefficiency. In addition, if Averch and Johnson allowed for $G^*_{x_1} < 0$, then

$$r_1/r_2 > \frac{G^*_{x_1}}{G^*_{x_2}} = -\frac{dx_2}{dx_1}$$

and the A-J Effect can be seen without any discussion of λ's. Now from equation (8) we have

$$(14)\quad r_1 = \alpha G^*_{x_1} + \lambda s_1 = G^*_{x_1} + \lambda\left(s_1 - G_{x_1}\right)$$

Thus, from equations (13) and (14)

$$r_1 > G^*_{x_1} \text{ if, and only if, } \alpha > 0, \lambda > 0$$

and

$$(15)\quad r_1 < G^*_{x_1} \text{ if, and only if, } \alpha > 0, \lambda < 0$$

However, the assumption $r_1 < s_1$ alone does not indicate the region of λ. Imposing the second assumption, i.e., concave revenue functions, we assume that G_{x_1} and G_{x_2} are continuous functions of x_1 and x_2, respectively, and that G_{x_1} and G_{x_2} can be represented by downward sloping curves as in Figure 1. The solution to the unconstrained profit maximization problem is given by

$$(16)\quad r_1 = G^0_{x_1}$$

with optimum solution x_1^0. For the constrained problem, equation (9) yields $G^*_{x_2} = r_2$, the same as for the unconstrained problem. Thus, from our previous discussion where we assumed that $x_1^* > x_1^0$ is the solution, we see that this assumption leads to

$$(17)\quad G^*_{x_1} < r_1 < s_1$$

From equation (17), (13), and (15) we would then conclude that $\alpha > 0$ and $\lambda > 0$ which implies $0 \leq \lambda < 1$. If, however, we assume no relationship between x_1^* and x_1^0, then we have no indication as to the position of $G^*_{x_1}$ relative to r_1 and s_1 and thus do not know the region of λ. Thus, a proof for the range of λ between 0 and 1 which depends on the continuity of G_{x_1} with x_1 must already assume a relationship between x_1^* and x_1^0.

We see, therefore, that the the A-J Effect is valid only if one accepts the two basic assumptions and one assumes either that $x_1^* > x_1^0$, or that $0 < \lambda < 1$. To prove the A-J Effect without either of these latter assumptions is impossible since the continuity of $\lambda = 0$ to $\lambda > 0$ cannot be shown in general.

REFERENCES

H. Averch and L. L. Johnson, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1052–69.

W. J. Baumol and A. K. Klevorick, "Input Choices and Rate of Return Regulation: Overview of the Discussion," *Bell J. Econ.*, autumn 1970, *1*, 162–90.

A. Takayama, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.*, June 1969, *59*, 255–60.

E. E. Zajac, "A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model," *Amer. Econ. Rev.*, Mar. 1970, *60*, 117–25.

Spectral Analysis and the Detection of Lead-Lag Relations

By JOHN C. HAUSE*

The difficulties of determining timing relationships between aggregate economic series by direct inspection of the series, by the cross correlation function, or by the formulation of explicit models are notorious. Empirical efforts to establish such relationships frequently fail or are highly inconclusive. In the search for techniques to deal with these problems, some economists have concluded that spectral and cross spectral analysis can avoid these difficulties and can provide direct and relevant information about leads and/or lags between pairs of economic time-series. A casual reading of studies such as those by Vittorio Bonomo and Charles Schotta, and T. J. Sargent might lead to the (incorrect) belief that it is generally possible to infer leads and lags by a simple transformation of the phase from cross spectral estimates.[1] This misunderstanding is found in a recent paper by Thomas Cargill that attempts to test the hypothesis that wage changes lag significantly behind price changes solely by an examination of cross spectral estimates.

This paper demonstrates that the interpretation of phase statistics depends critically upon the model (or class of models) that one assumes governs the relationship between a pair of time-series. Only under extremely restrictive assumptions is it valid to interpret phase information in the time domain in the way that Cargill proposes.

The discussion falls under two headings. The first deals with deterministic linear systems, distributed lags, and the interpretation of phase. The second section considers stochastic linear systems, the cross spectrum, leads and lags, and some final conclusions. The paper does not discuss statistical estimation.

I. Deterministic Linear Systems, Distributed Lags, and the Interpretation of Phase[2]

Much of the terminology of spectral analysis, *including the concepts of lead and lag*, originated in the analysis of deterministic systems (and their mathematical representations) by engineers. The engineering definitions of lead and lag are not closely related to the intuitive notion that most economists associate with these terms. The engineering concepts were originally used to describe phase relations between variables in the frequency domain. The engineering term corresponding most closely to the economist's notion of a lag in the time domain is "pure delay" (or simply delay.) Confusion over this terminology has doubtlessly been the main source of misunderstanding by economists of the interpretation of phase statistics. A brief analysis of deterministic linear models and the corresponding terminology clarifies these issues and provides the necessary background for interpreting spectral phase statistics.

Consider the following simple model of a stationary dynamic linear system $y(t) = Lx(t)$, in which $x(t)$ and $y(t)$ are variables (functions of time) representing input and output, respectively. A linear operator, L, characterizes the way in which the input is transformed into output.[3] It is generally

* Associate professor of economics at the University of Minnesota, and currently a research fellow at the National Bureau of Economic Research.

[1] These papers use the procedure described later in the text for converting phase shifts to the time domain without discussing the conditions under which this transformation is theoretically meaningful.

[2] Many detailed discussions of deterministic linear systems are available. An excellent, compact discussion is included in chapter 2 of G. M. Jenkins and D. G. Watts. For a more expanded treatment, see W. Kaplan.

[3] In the context of this model, an operator is linear if it transforms simultaneously applied inputs into output according to the relation

$$L(a_1x_1(t) + a_2x_2(t)) = a_1Lx_1(t) + a_2Lx_2(t)$$

where a_1 and a_2 are constants. The operator is "stationary" (time-invariant) if $Lx(t) = y(t)$ implies $Lx(t+t_0) = y(t+t_0)$, i.e., a translation of the input function in time by t_0 units implies the output is translated by the same amount of time. See Jenkins and Watts

possible to represent a linear dynamic input-output relationship as a distributed lag:

$$(1) \qquad y(t) = \int_0^{\infty} w(\tau)x(t - \tau)d\tau,$$

where $\omega(\tau)$ is a distributed lag function.[4]

An important characteristic of this class of models (for stable systems) is that if the input variable is a sinusoid of a specified frequency f, e.g., if $x(t) = \cos(2\pi ft)$, the output will be a sinusoid of the same frequency of the form

$$y(t) = G(f) \cos (2\pi ft + \alpha(f))$$

once the "transients" die out, where $\alpha(f)$ and $G(f)$ are real functions of f,[5] $\alpha(f)$ is the *phase shift* and is usually assigned a value between $-\pi$ and $+\pi$ (radians) by convention.[6] According to this convention, one says that at a given frequency f output *leads* input if α is positive, and that output *lags* input if α is negative. Thus the original definitions of lead and lag in the analysis of this class of models were simply a way of describing the shift in phase of the output sinusoid (*on the frequency domain, not the time domain*) and depend on a purely conventional way of measuring the size and sign of this phase shift.

The distributed lag formulation of this input-output model makes it clear that the output sinusoid is not in general obtained by a simple displacement of the input sinusoid in time by the phase angle α even though a graph of the input and output functions might lead one's intuition to this erroneous conclusion. The output is the superposition of the previous values of the input, with weights given by the distributed lag function $w(\tau)$. In general, $\alpha(f)$ will vary with frequency of the input even though the distributed lag itself is invariant to changes in frequency.

The concept of delay corresponds to the shift in time which many economists have in in mind when they discuss time lags between economic variables. A (linear) delay between an input and output may be expressed by the equation

$$y(t) = ax(t - t_0)$$

where a is a constant and t_0 is the length of the time delay. This is the only deterministic linear system for which there is a simple correspondence between the phase lead or lag and delay in real time. If the input is $x(t) = \cos(2\pi ft)$ and the system is a pure delay operator, the output

$$y(t) = a \cos (2\pi f[t - t_0])$$

Hence the phase $\alpha(f) = -2\pi ft_0$.[7] Dividing α by the frequency (in radians) of the input

(p. 36) or Kaplan for further discussion of stationary linear systems. Linear difference equations and the most widely used distributed lags in economics are examples of such systems.

[4] For discrete time models used in most econometric work, this expression is usually written in the form $y_t = \sum_{k=0}^{\infty} w_k x_{t-k}$, where at least one w_k differs from zero.

[5] $G(f)$ is the *gain* of the system, a nonnegative function that measures the amplitude of the output sinusoid if the amplitude of the input sinusoid is one.

[6] There are two conventions in use that should be distinguished, since the terminology and conventions discussed in this section are also relevant for cross spectral statistics. For the one described in the text (in which $-\pi \leq \alpha < \pi$; $G(f) \geq 0$), $G(f)$ is called the (cross) amplitude as well as the gain. The alternative convention restricts the phase range so that $-\pi/2 \leq \alpha < \pi/2$, and allows $G(f)$ to be positive or negative. (The source of these alternatives stems from the square root and the arctangent being multivalued functions.) The definitions of lead or lag (which depend on the conventional sign of the phase angle) are the same for these alternatives if the gain is positive according to both conventions; but they are reversed if the gain is negative according to the second convention.

Unfortunately, these conventions are not discussed adequately in many treatments of spectral analysis, including Jenkins and Watts, and C. W. J. Granger and M. Hatanaka. A computer program for estimating the cross spectral statistics based on such references might well adopt a bastardized (and for some purposes, misleading) convention with $F(f) \geq 0$, and $-\pi/2 \leq \alpha < \pi/2$. It would be highly advisable for economists using cross spectral computer programs to make certain which convention is being adopted in the computation of phase and gain statistics.

[7] Even if the pure delay model is appropriate, the convention that determines α, as discussed in fn. 6, must be handled with care. For the convention where $-\pi \leq \alpha < \pi$, if f is greater than $1/2t_0$, $-2\pi ft_0$ will fall outside the conventional range. In this case, an integral multiple of π must be added to the conventionally measured phase to obtain the total phase angle delay $2\pi ft_0$. This problem is discussed further in Granger and Hatanaka.

sinusoid gives the length of the delay, since $\alpha/2\pi f = -t_0$. For all other linear input-output systems, it is incorrect to interpret $\alpha/2\pi f$ as if there is a pure delay between input and output.

One further important characteristic of deterministic linear systems useful in the next section is cited without proof. If the system is represented in distributed lag form, the Fourier transform of the distributed lag of $w(t)$ is

$$\int_{-\infty}^{\infty} e^{-i2\pi ft} w(t)dt = G(f)e^{i\alpha(f)}$$

where $\alpha(f)$ and $G(f)$ are the phase and gain as defined above. This formula shows that the gain and phase of a linear system can be obtained directly from the Fourier transform of the distributed lag of the system.

II. Stochastic Linear Systems, the Cross Spectrum, and Leads and Lags[3]

It can be shown that the cross spectrum of two stationary random processes, x and y, is identical to the Fourier transform of the cross correlation function of x and y, and that the cross spectrum can be written in the form

$$A_{xy}(f)e^{i\beta_{xy}(f)}$$

where $A_{xy}(f)$ and $\beta_{xy}(f)$ are real functions of f called the cross amplitude spectrum and phase spectrum, respectively.

Suppose that some stochastic elements are introduced into the simple linear input-output model discussed in Section I. Using distributed lag notation, let output

$$(2) \qquad y(t) = \int_{0}^{\infty} w(\tau)x(t-\tau)d\tau + u(t),$$

where $u(t)$ is an additive random component. Assume $u(t)$ is statistically independent of $x(t)$. The cross spectrum of input and output of this system is $C_{xy}(f) = W(f)C_{xx}(f)$, i.e., the cross spectrum is the product of the Fourier transform of the distributed lag function and the spectrum of x. The spectrum must be a real nonnegative function. The preceding section pointed out that $W(f)$ can be written in the form $G(f)e^{i\alpha(f)}$. Thus for this linear model

$$A_{xy}(f) = [G(f)C_{xx}(f)] \quad \text{and} \quad \beta_{xy}(f) = \alpha(f)$$

The last equation states that the phase angle of the (deterministic) distributed lag in this case is identical to the phase of the cross spectrum of input and output. The result is not surprising, since spectral analysis essentially amounts to a frequency decomposition of time series. For each frequency, the phase relationship of input and output is precisely that determined by the distributed lag.

The discussion in Section I emphasized that the phase α of a deterministic system can be interpreted as a shift in time only if the distributed lag is a pure delay. The calculations and conclusions in Cargill's paper are based on the assumption that phase has been determined in a pure delay system. The variable *tau* is defined by the relation $\alpha(f)/2\pi f$, and *tau* is discussed as if it measures a pure delay at that frequency. A simple, but revealing illustration in the Appendix demonstrates how misleading this procedure can be if the distributed lag is not a pure delay.

The emphasis upon the analysis of pairs of variables in which one variable can be regarded as a distributed lag of the other stems from the important role this model has played in thinking about the time relationships between certain aggregate economic variables. E. Malinvaud (p. 473) has given particular emphasis to the role of distributed lags in econometric investigations and has suggested why they are more plausible than pure delays. If one turns to linear models which are not simple input-output systems, the interpretation of phase (and *tau*) becomes even more complicated. For example, the pair of variables used for computing cross spectral statistics might themselves be distributed lags of a third variable. The nature of the cross spectrum of such a pair of variables is readily determined, and the calculation shows that no simple interpretation of delay can be associated in general with this model. Another possible

[3] Jenkins and Watts (ch. 8) and Granger and Hatanaka (ch. 5) contain a more detailed discussion of cross spectra.

TABLE 1—PHASE ANGLE AND *TAU*

Implied by Exponentially Declining Weights Distributed Lag ($\beta = .7$)

$2\pi f$ (radians)	Period (1/f)	Phase $-\alpha$ (radians)	$-Tau$ (time units of the sampling) period)
.1	62.8	.226	2.26
.5	12.5	.716	1.43
.9	7.0	.771	.86
1.7	3.7	.567	.33
2.5	2.5	.262	.10

linear model arises when there is feedback between the variables. This model and the explicit cross spectral statistics that are determined by it have been discussed by C. W. J. Granger and by the present author. Here again there are generally no simple conclusions that can be derived about the time domain relationships of the variables solely on the basis of the phase shift.

An intuitive suggestion has been made by Granger and M. Hatanaka that in some cases one might regard low frequencies as reflecting "long-run" relationships, while high frequencies correspond to the "short-run." Suppose that the spectral decomposition of the time-series statistically isolates essentially independent factors governing long, intermediate, and short runs. Finally assume that for each "run" the relationship between the economic variables is a pure delay. Under these conditions, the computation of *tau* for the frequencies corresponding to different runs yield the pure delay for that run. The procedure adopted by Cargill seems to rest implicitly on some argument similar to the one just sketched. But this model seems very implausible for the wage and price variables on which his study is based.

While such a model is conceivable, there is a large burden of proof in establishing its relevance for a particular pair of economic series. The assumption that each frequency band with high coherence corresponds to an independent pure delay surely requires an explicit justification. One can imagine time-series containing strong periodicities from seasonal factors, where the seasonal factors may be largely independent from other factors. Even here, some analysis is required to explain why a pure delay between the variables at the seasonal frequency is more plausible than a distributed lag.

The main conclusion of this analysis is that phase leads and lags measured from cross spectral estimates will rarely provide economists with direct estimates of the time domain relationships that are of interest. To avoid terminological confusion, it might be useful for economists using spectral techniques to reserve the unmodified words "lead" and "lag" exclusively for descriptions of phase relations, and to describe translation in the time domain by the phrase pure delay. If it is assumed in some application that phase shift or *tau* corresponds to pure delay, evidence should be provided to justify the assumption. If phase information is not intended to convey the impression of pure delay, an explicit warning is desirable. Finally, the discussion of all cross spectral analyses would be greatly improved if an explicit model (or class of models) is presented of the assumed dynamic relationship linking the variables.

APPENDIX

Calculations of Tau for a Simple Input-Output Distributed Lag System

Suppose that input and output are related by the discrete time distributed lag of exponentially declining weights that often appears in econometric models: $w_k = \beta^k$ $(k = 0, 1, \ldots)$. In this case

$$(3)\quad W(f) = \sum_{k=0}^{\infty} \beta^k e^{-i2\pi fk} = 1/(1 - \beta e^{-i2\pi f})$$

From this complex function one obtains the following formula for the phase:

$$(3')\quad \alpha(f) = -\arctan[(\beta \sin 2\pi f)/(1 - \beta \cos 2\pi f)]$$

Dividing this expression by $2\pi f$ gives the time lag, *tau*, in Cargill's terminology.

Table 1 shows that for the simple exponential lag, $\alpha(f)$ is always negative. As frequency increases, the phase angle initially increases in magnitude until $\cos 2\pi f = \beta$, and thereafter declines. *Tau* itself is a declining function of f.

These results demonstrate that it is highly misleading to interpret *tau* as the lag between economic series linked by this input-output model. The time response of the system to changes in the input is given by the exponentially declining weights, and not by the phase angle or *tau*.

REFERENCES

V. Bonomo and. C. Schotta, "A Spectral Analysis of Post-Accord Federal Open Market Operations," *Amer. Econ. Rev.*, Mar. 1969, *59*, 50–61.

T. F. Cargill, "An Empirical Investigation of the Wage-Lag Hypothesis," *Amer. Econ. Rev.*, Dec. 1969, *59*, 806–16.

C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica*, July 1969, *37*, 424–38.

——— and M. Hatanaka, *Spectral Analysis of Economic Time Series*, Princeton 1964.

J. C. Hause, "Leads, Lags and Spectral Analysis," *Econometrica*, Oct. 1964, *32*, 687.

G. M. Jenkins and D. G. Watts, *Spectral Analysis and its Applications*, San Francisco 1968.

W. Kaplan, *Operational Methods for Linear Systems*, Reading, Mass. 1962.

E. Malinvaud, *Statistical Methods of Econometrics*, Chicago 1966.

T. J. Sargent, "Interest Rates in the Nineteen-Fifties," *Rev. Econ. Statist.*, May 1968, *50*, 164–72.

Subsidized Housing in a Competitive Market: Comment

By GORDON TULLOCK*

Edgar Olsen's recent article in this *Review* is a significant contribution to clarifying the economics of a complex and difficult area. It is not the purpose of this comment to raise any questions as to his economic analysis, but to point out that, granted competition, it might be extremely difficult to subsidize low-income family consumption of superior housing by the method he suggests. In a sense, my objection is against interest, since I myself would much prefer that any subsidies on housing for lower income families use the Olsen method rather than the method of direct government provision. As I shall suggest later, however, there is another possible procedure which I regard as superior to either the provision of public housing or the Olsen subsidy.

For simplicity, assume that some poor person receiving subsidies under the Olsen procedure would normally spend $60 a month on rent. He is permitted to purchase for $60 a $100 rent certificate which can then be used to rent superior housing. This amounts to giving him an income supplement of $40, but attempting to compel him to use it for one particular purpose. His utility would be higher if you simply gave him the $40 and permitted him to spend it on anything he chose. It seems reasonable if you did so he would indeed improve his housing, but would also improve his consumption of other goods as well. To use a rough rule of thumb, let us assume that if he were given $40 in cash every month, he would choose to spend $10 of this in increasing his consumption of housing—renting an apartment at $70—and spend the other $30 on other matters. Clearly, from the standpoint of the poor person the receipt of a direct subsidy would be superior.

Granted that this is so and that the market is highly competitive (even if not perfectly competitive), it seems likely that the individual would be able to find a landlord who is willing to rent him an apartment which is normally worth $70 for the $100 certificate, and then make an under-the-table rebate to him of $30. Olsen says that, "It would be illegal to exchange these certificates for other than housing services," but it seems to me that this is a type of crime which is extremely hard to detect. The only people involved would be the landlord and the poor tenant, and both would benefit from the crime. The so-called "crimes without victims," such as gambling, prostitution, drug sale, all present problems in enforcing the law. In fact many people, including myself, feel that these laws should be repealed. Olsen, in effect, is creating a new "crime without victim," and we can assume similar problems in enforcement. Further, intensive police activity to limit rebates might create a risk with the result that both the poor person and the landlord would be worse off than if the police activity did not exist, but most rebating would continue.

From the standpoint of the recipient, such a rent certificate is inferior to a direct cash payment. Surely there is some cash payment of slightly less cost to the state than the rent certificate which would be, from the standpoint of the recipient, superior to the certificate. Granting this, it seems to me that we should aim at direct cash payment. I presume that Olsen would agree. The reason he is advocating this particular mechanism is because he believes that subsidies aimed at increasing poor persons' consumption of housing services, rather than simply increas-

* Center for Study of Public Choice, Virginia Polytechnic Institute.

ing their income, are a more or less permanent part of our economy, and he wishes to make them more efficient. I cannot quarrel with this desire on his part, but I doubt that his particular technique would work. It seems to me that we would be better advised to try to change government policy toward raising the incomes of the poor rather than trying to adjust their consumption toward the qualitative standards of those who are not poor.

REFERENCE

E. A. Olsen, "A Competitive Theory of the Housing Market," *Amer. Econ. Rev.*, Sept. 1969, *59*, 612–22.

Subsidized Housing in a Competitive Market: Reply

By EDGAR O. OLSEN*

Gordon Tullock has called to our attention the difficulty in preventing the recipient of a rent certificate from converting his certificate into the equivalent of a cash grant. His comments apply not only to certificates for housing but also to certificates for other goods, e.g., food stamps. Even though phrased in terms of housing, my reply has the same generality.

In the framework of Paretian welfare economics, the cost of enforcing the requirement that rent certificates be spent only on housing is relevant to determining the best means for redistributing consumption. Hence, Tullock *may* be correct in saying that this cost is so large that all transfers should be in the form of cash. However, this is not necessarily the case, as I demonstrate in the second part of my reply.

It is necessary to take exception to a belief explicit in Tullock's comment. He clearly argues that transfers in kind would result in inefficient resource allocation even if it were costless to prevent recipients from converting their in-kind subsidy into cash. He attempts to prove his point by showing that the recipient and his landlord could benefit by violating the provision of the certificate, and that no third party would be hurt by this violation.[1] He suggests that I advocate rent certificates because I believe that housing subsidies, though undesirable, are inevitable and because I want the housing service consumed by subsidized families to be produced efficiently, as it would be by private producers in a competitive market. In fact, I think that housing subsidies and certain other in-kind transfers are desirable provided that their recipients can be costlessly prevented from converting the in-kind transfer into a cash grant.

In the first part of this reply, I will show that there exists a set of indifference maps and a societal budget constraint such that a costlessly enforced rent certificate scheme will result in efficient resource allocation. Allowing the recipient to violate the restriction on the use of his certificate will make someone worse off and will result in inefficient resource allocation. Existence is proven by means of an example in which there is a consumption externality. In this example, an infinite number of Pareto optimal allocations may be attained by rent certificate plans but not by unrestricted cash grants. One of these optima is the Lindahl solution associated with the pre-transfer distribution of income.[2] It is because I believe that consumption externalities are pervasive and because I attach special normative significance to the Lindahl solution that I proposed rent certificates. However, the case for rent certificates is stronger than this at least in my example because all Pareto optimal allocations at which all people are better off than they would be in the absence of transfers may be attained by rent certificates but not by unrestricted cash grants.

The assumptions and results of this example will be stated in the paper; a guide to

* Assistant professor of economics, University of Virginia. I am especially grateful to Stanley M. Besen, Edgar K. Browning, Joseph S. DeSalvo, Harold M. Hochman, Roland N. McKean, James D. Rodgers, and Gordon Tullock for comments which were helpful in revising an earlier draft of this paper.

[1] A similar argument attempting to show that it is better to give recipients unrestricted cash grants than to reduce the price that they pay for one good is due to Alan T. Peacock and D. Berry. Their argument about negative income and excise taxes follows easily from an earlier discussion of positive income and excise taxes which is reproduced and criticized by Milton Friedman. My objection to Tullock's proof is different from Friedman's.

[2] As Paul Samuelson (1969, p. 102) makes clear, a good generating a consumption externality is a public good by his definition. For example, if A is concerned about B's housing, then housing consumed by B is a public good. See Richard Musgrave (pp. 73–78) for the Lindahl solution to the determination of the optimal quantity of a public good and the optimal distribution of taxes to pay for it.

the mathematical derivation of these results may be obtained from the author.

I. Optimality with Costless Enforcement

Assume that there are two individuals in society, the grantor and the recipient. They consume two goods, nonhousing X and housing H. The grantor directly consumes only nonhousing, but he also cares about the recipient's consumption of housing. The recipient directly consumes both nonhousing and housing. Let X_g and X_r be the quantities of nonhousing directly consumed by the grantor and recipient, respectively, and H_r be the quantity of housing directly consumed by the recipient. Suppose that the indifference maps of the grantor and the recipient are

$$X_g^{.9}H_r^{.1} = a \tag{1}$$

$$X_r^{.8}H_r^{.2} = b \tag{2}$$

where X_g, X_r, and H_r are nonnegative. Assume that nonhousing and housing are produced at constant costs of \$2 and \$1 per unit, respectively. Finally, assume that the grantor has an income of \$400 and the recipient an income of \$100 per time period.[3] Hence, the budget frontier of this society is

$$2(X_g + X_r) + H_r = 500 \tag{3}$$

In this example there are infinitely many Pareto optimal consumption patterns within society's budget constraint. Samuelson (1954, 1955) would choose among these patterns by means of a social welfare function depending only on the utility indices of the members of society. This is not the only means of selecting one of the Pareto optimal allocations to be the grand optimum. Indeed, my belief in the desirability of rent certificates is based partly on a normative theory in the tradition of Lindahl in which the grand optimum is determined by another means. This theory is based upon individual preferences, the pretransfer distribution of income, and certain pricing rules.[4]

The normative pricing rules of the theory are as follows. If the good is a private good, then the theory says that each consumer should pay the marginal cost of producing it. In my example, nonhousing is a private good. Hence, both the grantor and the recipient should pay \$2 for each unit of nonhousing that they consume. If a good is a public good, then each person should pay a price for it such that the marginal rate of substitution between the public good and the private good is equal to the ratio of his price of the public good to the marginal cost of producing the private good. In my example, the mathematical representations of these rules are

$$\frac{\partial(X_g^{.9}H_r^{.1})/\partial H_r}{\partial(X_g^{.9}H_r^{.1})/\partial X_g} = \frac{P_g^h}{2} \tag{4}$$

$$\frac{\partial(X_r^{.8}H_r^{.2})/\partial H_r}{\partial(X_r^{.8}H_r^{.2})/\partial X_r} = \frac{P_r^h}{2} \tag{5}$$

where P_g^h is the price to be paid by the grantor and P_r^h is the price to be paid by the recipient for each unit of housing that the recipient consumes. These normative prices are the prices which determine the optimal taxes according to the benefit approach of Lindahl and Bowen. Equations (4) and (5) are equivalent to

$$2X_g - 9P_g^hH_r = 0 \tag{6}$$

$$X_r - 2P_r^hH_r = 0 \tag{7}$$

Of course, the sum of the prices paid by different people for each public good must be equal to the marginal cost of producing it if, as assumed, the marginal cost is constant.

[3] To keep the exposition simple, I work with an exchange model and abstract from production. We might think of this society as composed of two retired people living on annuities and buying both goods in the perfectly competitive markets of another society. The production side could be considered in a trivial way by assuming that there is one factor of production, that one unit of nonhousing can be produced by two units of the input and one unit of housing by one unit of the input, and that the grantor owns 400 units and the recipient 100 units of the productive factor. In this case, equation (3) would be a production possibility frontier.

[4] See Olsen (1969) for a recent elaboration of this theory.

In my example,

$$P_g^h + P_r^h = 1 \quad (8)$$

Finally, this normative theory assumes that each individual should consume only as much of the goods as can be bought with his initial income at the optimal prices. Therefore,

$$2X_g + P_g^h H_r = 400 \quad (9)$$

$$2X_r + P_r^h H_r = 100 \quad (10)$$

Equations (6) through (10) are the mathematical representation of the normative theory in this particular example. Therefore, we have five equations, only one of which is linear, and five unknowns, X_g, X_r, H_r, P_g^h, and P_r^h. Despite the nonlinearities there is only one solution ($X_g = 180$, $X_r = 40$, $H_r = 60$, $P_g^h = 2/3$, $P_r^h = 1/3$) to this system of equations. Hence my grand optimum allocation of resources is ($X_g = 180$, $X_r = 40$, $H_r = 60$). The utility indices of the grantor and the recipient at this allocation are about 161.3 and 43.4. It can be shown that this allocation of resources is one of the infinite number of Pareto optimal allocations given the indifference maps (1) and (2) and the societal budget constraint (3). It can also be proven that this allocation cannot be reached by cash grants alone. Therefore, subsidies in kind do not necessarily result in inefficiency.

In the absence of transfers between the grantor and the recipient, the grantor would spend all of his income on nonhousing. Hence, he would consume 200 units of this good. The recipient would choose to consume 40 units of nonhousing and 20 units of housing. The utility indices of the grantor and the recipient at this allocation are about 158.9 and 34.8. It can be shown that this allocation of resources is not Pareto optimal. Thus, there are circumstances in which transfers are necessary for efficient resource allocation.[5]

In this situation I would propose that the government sell to the recipient a rent certificate with a face value of \$60 and charge the recipient \$20. At this point in the analysis, I assume that the government can costlessly force the recipient to use his certificate for housing only, and I propose that the government do so. Though the recipient is free to spend more than \$60 on housing, he will not choose to do so in this case. Hence, my proposals would result in the recipient's consuming 60 units of housing and 40[=(\$100−\$20)/\$2] units of nonhousing. I propose that the government collect \$40 in taxes from the grantor. The grantor would buy 180 units of nonhousing with the \$360 that he would have left. With \$20 from the recipient and \$40 from the grantor, the government redeems the certificate from the seller of housing at face value. These proposals would result in my grand optimal allocation of resources. This example proves that there are situations in which rent certificates will result in efficient resource allocation.[6]

Suppose that we allowed the recipient to redeem his rent certificate at face value for cash but prohibited further transfers. The grantor would continue to consume 180 units of nonhousing. The recipient will exchange his certificate for \$60 and will have \$140 [=\$100−\$20+\$60] in cash to spend as he pleases. In this case, he would consume 56 units of nonhousing and 28 units of housing. The utility indices of the grantor and the recipient at this allocation are about 149.4 and 48.8. Naturally, the recipient prefers this allocation of resources to my grand optimum. However, the grantor prefers my grand optimum. Indeed, in this particular example he prefers the allocation in the absence of transfers to this allocation. Since it is this type oi model which leads me to recommend rent certificates, I cannot agree with Tullock that my restriction on the use

[5] This has already been proven in more general cases by Otto Davis and Andrew Whinston and by Harold Hochman and James Rodgers.

[6] This is not to say that rent certificates are necessary for efficient resource allocation but only that they are sufficient. My grand optimum could be attained by lowering the price per unit of housing to the recipient to one third of a dollar and allowing him to consume any quantity that he chooses. Furthermore, there is one Pareto optimal allocation in this example that can be attained by an unrestricted cash transfer.

of the rent certificate creates a crime without a victim. The grantor is the victim. Furthermore, the allocation ($X_g=180$, $X_r=56$, $H_r=28$) is not Pareto optimal. Both the recipient and the grantor would be better off, for example, at the feasible consumption pattern ($X_g=170$, $X_r=440/9$, $H_r=560/9$). Therefore, to allow recipients of subsidies in kind to convert these subsidies into cash may result in inefficient resource allocation.

In this example there is only one Pareto optimum ($X_g=0$, $X_r=200$, $H_r=100$) that can be attained by a cash transfer between the grantor and the recipient.[7] Therefore, even in the presence of consumption externalities, subsidies in kind are not necessary for efficient resource allocation. Since the grantor would consume nothing directly in this situation, he would be as bad off as he could be (i.e., his utility index would be zero). Under these circumstances, I doubt that anyone other than the recipient would argue for unrestricted cash grants.

Attainment of the infinity of other Pareto optimal allocations requires noncash grants. An infinite subset of this infinite set of Pareto optima is composed of allocations which both the recipient and the grantor prefer to the allocation in the absence of transfers. Each of these allocations can be attained by a voluntary rent certificate scheme. In a forthcoming paper, I have proved that if the recipient prefers one of the consumption patterns which the rent certificate scheme makes available to him to his pretransfer pattern, then either the rent certificate is equivalent to an unrestricted cash grant or the recipient will spend precisely the face value of his certificate on housing. Since the infinite subset consists only of allocations not attainable by cash transfers, the second possibility obtains. Therefore, if the allocation (X_g^*, X_r^*, H_r^*) is a Pareto optimal allocation preferred by both the grantor and the recipient to the pretransfer allocation of resources, then this allocation can be attained by offering to sell to the recipient a rent certificate with a face value of $\$H_r^*$ for $\$100-\$2X_r^*$ and taxing the grantor $\$H_r^*-\$100+\$2X_r^*$.

[7] In the many goods, many persons case, there will be one such point for each person who is concerned only about goods that he consumes directly because the allocation of resources that would result from transferring all of society's income to such a person is Pareto optimal.

II. Optimality with Costly Enforcement

Throughout the previous section, I assumed that the recipient could be forced costlessly to spend the face value of his rent certificate on housing. I now dispense with this assumption.

We know that the recipient would prefer to exchange his certificate for its face value in cash. Therefore, unless some attempt is made to enforce the provision on the use of his rent certificate, he will violate it. In order to deter the recipient from violating this provision, there must be some positive probability of being caught. To create this probability some resources must be expended on law enforcement. If resources are expended on law enforcement, then the allocation ($X_g=180$, $X_r=40$, $H_r=60$) is not attainable because some of the money spent on nonhousing and housing must be diverted to law enforcement.

In my example the violation of the restriction on the use of the rent certificate is a crime with a victim. When should this crime be prevented?[8] On grounds of efficient resource allocation, a crime should be prevented if, and only if, the loss to the victim exceeds the gain to the criminal by more than the minimum cost of preventing the crime. In my example, the loss to the grantor due to the recipient's violation would be $29.24. That is, the grantor is willing to pay up to this amount to prevent the recipient from violating the restriction by exchanging his certificate for cash at face value. The gain to the recipient from this violation would be $15.43. That is, if the restriction were not enforced, then the recipient would be as well off as he would have been had he been given an unrestricted cash grant of $40; if the restriction were enforced, then the recipient would be as well off as he would have

[8] Gary Becker and Harold Demsetz have investigated this general question in some detail.

been with a cash grant of \$24.57. Therefore, if the cost of enforcing this provision is more than \$13.81[=\$29.24−\$15.43], then on grounds of efficient resource allocation the provision should not be enforced. In this case I would agree with Tullock that the recipient should be given a cash grant. However, if the cost of enforcement is less than \$13.81, then the provision should be enforced. In this case, I would say that an enforced rent certificate plan is preferable to a cash grant.

III. Concluding Comments

Having shown that rent certificates can be justified in the framework of Paretian welfare economics and that there is some cost of enforcing the restriction on the use of the certificates so low that rent certificates are preferable to cash grants, I now wish to agree with Tullock that the cost of enforcing this provision is likely to be very high. As a result, I think that it is entirely reasonable to argue for transfers in cash even though consumption externalities are pervasive. This, however, is an empirical question. It cannot be settled solely on theoretical grounds.

I proposed rent certificates because I believe that there are many paternalistic altruists in this country and that housing is one of the goods that these people think the poor value too lightly.[9] My belief stems from the casual observation that most governmental and nongovernmental transfers to the poor are in kind (e.g., public housing, food stamps, and medicare). If this sort of consumption externality proves to be unimportant, then I will withdraw my rent certificate proposal.

[9] A discussion of other consumption externalities that have been postulated to justify housing subsidies appears in my dissertation (pp. 19–21, 38–48).

REFERENCES

G. S. Becker, "Crime and Punishment: An Economic Approach," *J. Polit. Econ.*, Mar./Apr. 1968, *76*, 169–217.

O. A. Davis and A. B. Whinston, "Welfare Economics and the Theory of the Second Best," *Rev. Econ. Stud.*, Jan. 1965, *32*, 1–14.

H. Demsetz, "The Exchange and Enforcement of Property Rights," *J. Law Econ.*, Oct. 1964, *7*, 11–26.

M. Friedman, "The 'Welfare' Effects of an Income Tax and an Excise Tax," *J. Polit. Econ.*, Feb. 1952, *60*, 25–33; reprinted in M. Friedman, *Essays in Positive Economics*, Chicago 1953, 100–113.

H. M. Hochman and J. D. Rodgers, "Pareto Optimal Redistribution," *Amer. Econ. Rev.*, Sept. 1969, *59*, 542–57.

R. A. Musgrave, *The Theory of Public Finance*, New York 1959.

E. O. Olsen, "A Welfare Economic Evaluation of Public Housing," unpublished doctoral dissertation, Rice Univ. 1968.

———, "A Normative Theory of Transfers," *Publ. Choice*, spring 1969, *6*, 39–58.

———, "Some Theorems in the Theory of Efficient Transfers," *J. Polit. Econ.*, forthcoming.

A. T. Peacock and D. Berry, "A Note on the Theory of Income Redistribution," *Economica*, Feb. 1951, *18*, 83–90.

P. A. Samuelson, "The Pure Theory of Public Expenditure," *Rev. Econ. Statist.*, Nov. 1954, *36*, 387–89.

———, "Diagrammatic Exposition of a Theory of Public Expenditure," *Rev. Econ. Statist.*, Nov. 1955, *37*, 350–56.

———, "Pure Theory of Public Expenditure and Taxation," in J. Margolis and H. Guitton, eds., *Public Economics*, New York 1969.

Expectations and the Demand for Bonds: Comment

By RICHARD ROLL*

John H. Wood concluded his recent article in this *Review* by stating

> . . . the expectations hypothesis [of the term structure of interest rates] is logically invalid . . . the Hicks and Lutz equations rely on sub-optimal decision rules and, consequently, are without behavioral significance. . . . The awkward Hicks and Lutz formulations have hindered inquiries into the effects of uncertainty on the term structure . . . our results have made it possible for the study of the structure of rates under such conditions to proceed on a sound theoretical basis. [pp. 529–30]

The purposes of this comment are to determine the historical accuracy of Wood's interpretation of Irving Fisher (pp. 273–74), J. R. Hicks (pp. 144–47), and F. A. Lutz (pp. 499–529) and to examine the economic validity of Wood's theory.

I. Fisher, Hicks, and Lutz

According to Wood, the traditional expectations hypothesis of the term structure implies a decision rule of the following form: A trader[1] should ". . . prefer n-period bonds, be indifferent between n- and one-period bonds, or prefer one-period bonds, when"

$$(1)\quad (1+R_n) \gtreqless \cdot[(1+R_1)(1+{}_1r_1)\dots(1+{}_{n-1}r_1)]^{1/n}$$

(p. 524.)

where R_k is the current observed market yield on k-period bonds and ${}_jr_1$ is the yield currently expected by the trader to prevail on one-period bonds j periods hence.

At first glance, expression (1) might seem an accurate algebraic representation of the expectations hypothesis. A one-period forward rate applicable j periods hence is, by definition,

$$(1+{}_{j-1}\rho_1) \equiv \frac{(1+R_j)^j}{(1+R_{j-1})^{j-1}}$$

so that a current market yield in terms of forward rates is

$$(1+R_n)\equiv[(1+R_1)(1+{}_1\rho_1)\dots(1+{}_{n-1}\rho_1)]^{1/n}$$

The "pure" expectations hypothesis states that each observed market forward rate is equal to its corresponding expected future spot rate, (see David Meiselman, p. 10) ${}_j\rho_k={}_jr_k$, so that the equality form of (1) is the market equilibrium equation.[2] Indeed, Wood indicates this fact and then destroys our belief in the logic of the expectations hypothesis by presenting an alternative decision rule which earns a higher expected return.

The trader using Wood's decision rule will ". . . prefer n-period bonds, be indifferent between n- and m-period bonds, or prefer m-period bonds, when"

$$(2)\quad (1+R_n)\gtreqless(1+R_m)^{m/n}=\frac{(1+{}_1r_{n-1})^{(n-1)/n}}{(1+{}_1r_{m-1})^{(m-1)/m}}$$

(p. 525.)

Decision rule (2) is based on a strategy of maximizing return one period at a time. In the first period, period 0, the investor should make pairwise comparisons of all bonds and select the bond whose expected holding period yield during the first period is highest. Wood proves that decision rule (2) is superior to (1). Rule (1) is sub-optimal because it prompts the trader, who is indifferent to risk, into basing current decisions partly on an-

* Carnegie-Mellon University

[1] We shall assume (with Wood) that this trader operates in a bond market free of transaction costs, maximizes return over a horizon as long or longer than the longest-term bond outstanding, and is indifferent to risk.

[2] This interpretation of (1) requires the r's to be rates expected by some hypothetical composite market trader.

ticipated future decisions. There is no doubt that (2) is a rule superior to (1). The question, however, is whether decision rule (1) can be ascribed to the expectations hypothesis.

An examination of Fisher, Hicks, and Lutz will show that no such decision rule was ever recommended by them. In fact, Fisher and Lutz presented the same example that Wood used to show the inferiority of rule (1). This example, to be discussed in detail in the next section, involves a trader who expects a rise in the rate on long-term bonds. As Wood shows, such a trader will prefer short-term bonds now (p. 527) and will wait until rates have risen to buy long-term bonds. Compare Fisher's analysis:

> Those who expect the [long-term] rate of interest to fall will prefer to invest in long-time securities at the present market rates, even when those rates are less than on securities of shorter time, *while those who expect the [long-term] rate of interest to rise will prefer short-time securities.* (italics added) [p. 274]

and Lutz':

> The second possibility is that the investor may expect the yield on the bond at some intermediate date to exceed the average of the short rates from that date onwards, i.e., he expects the market price of the bond to be relatively *low* at that date. He will then contemplate going into the short market now and into the long market later. [p. 514–15]

Fisher and Lutz were clearly aware that decision rule (1) leads to incorrect actions.

Hicks, the third author Wood associates with rule (1), should probably be left out of the discussion entirely. His theory of the term structure is intricately connected to risk which we have assumed away here.[3] Let it suffice to note that decision rule (1) cannot be found in his book.

Finally, we must mention that all three writers, Fisher, Hicks, and Lutz, were discussing the term structure as a market equilibrium phenomenon and not as a normative theory for the guidance of investors. It is quite easy to be misled by the equality form of (1), which is the market equilibrium condition, into accepting the inequalities, which have nothing whatever to do with the theory. A thorough reading of these early articles will demonstrate the unfairness of Wood's criticism, "Discussions of the expectations theory of the term structure of interest rates have tended to be rather mechanical, ignoring the microeconomic foundations of market equilibrium solutions" (p. 522). Nothing could be less true. All three authors owe the frequent references to their work to a *concentration* on the foundations of market equilibrium and a style of expression that is lucid and *non*-mechanical.

II. A Revised Bond Investor's Decision Rule

We now turn from Wood's interpretation of history to a discussion of his decision rule. This section intends to prove that neither Rule (1) *nor* Rule (2) is optimal.

The correct decision rule can be demonstrated with Wood's three-period example. He assumed that two risk-indifferent traders, G and H, held expectations depicted by[4]

$$\text{(H.1)}\quad (1+R_2)^2=(1+R_1)(1+{}_1r_1)$$

$$\text{(H.4)}\quad (1+R_3)^3>(1+R_1)(1+{}_1r_1)(1+{}_2r_1)$$

$$\text{(G.4)}\quad (1+R_3)^3<(1+R_1)(1+{}_1r_2)^2$$

The lower case r's denote the trader's expected future spot rates and upper case R's denote market spot rates at period 0. Trader G's inequality (G.4) was Wood's decision rule (2) and Trader H's inequality (H.4) was intended to be the decision rule (1) implied by the expectations hypothesis. Wood assumed, via (H.1), that both G and H were indifferent between one- and two-period bonds.[5] According to these decision rules, H prefers three-period over one-period bonds and G prefers the opposite. Wood showed

[3] Lutz was also concerned with risk and devoted much of his article to its discussion. Hicks, however, discussed practically nothing else.

[4] The H and G notations are Wood's.

[5] Both H and G were assumed to have made a decision to commit their funds for at least three periods.

that G's one-period gain would indeed be greater than H's. H is led into error because he "... expects at time 0 to prefer two-period over one-period securities at time 1. Because of his use of decision rule (2), he is influenced by this expected future preference in his portfolio decision at time 0" (p. 527).

Neither the rule Wood recommended (2) nor the rule he attributed to the expectations hypothesis (1) will lead to optimal actions by a risk-indifferent trader. The correct rule is (a) *make one-period spot loans now with available resources and* (b) *make forward loans now if the forward rate is greater than the corresponding expected future spot rate.* In the present example, the forward rates are

(3) $(1 + {}_1\rho_1) = (1 + R_2)^2/(1 + R_1)$

(4) $(1 + {}_2\rho_1) = (1 + R_3)^3/(1 + R_2)^2$

(5) $(1 + {}_1\rho_2)^2 = (1 + R_3)^3/(1 + R_1)$

where ${}_j\rho_k$ is the k-period forward rate to begin j periods hence.

To determine the optimum investment at time zero, the trader must compare forward rates to expected spot rates as follows: ${}_1\rho_1$ to ${}_1r_1$, ${}_1\rho_2$ to ${}_1r_2$, and ${}_2\rho_1$ to ${}_2r_1$.

Using equations (H.1) and (3), we obtain for the first comparison,

(6) $$1 + {}_1r_1 = 1 + {}_1\rho_1$$

which indicates that now, in period 0, the trader is indifferent between making one-period forward loans to begin one period hence and waiting until period 1 to make one-period spot loans.

Using the inequality (G.4) and equation (5), we have

$$(1 + {}_1\rho_2)^2 = \frac{(1 + R_3)^3}{(1 + R_1)} < (1 + {}_1r_2)^2$$

or

(7) $${}_1\rho_2 < {}_1r_2$$

Since the two-period forward rate to begin one period hence is less than the two-period expected spot rate, the trader should issue forward loans now. He should borrow forward. In a world of perfect capital markets and zero transaction costs, he can do this by selling three-period bonds short and buying one-period bonds with the proceeds. At the beginning of period 1, he will receive an expected capital gain of

(8) $$d_1\left[(1 + R_1) - \frac{(1 + R_3)^3}{(1 + {}_1r_2)^2}\right]$$

where d_1 is the dollar amount of three-period bonds sold short and one-period bonds purchased. By referring to inequality (G.4), one can verify that the expected gain, represented by (8), is indeed a positive quantity.

Using equations (H.1) and (4) and inequality (H.4), we obtain the third comparison,

$$(1 + {}_2\rho_1) = \frac{(1 + R_3)^3}{(1 + R_2)^2} > 1 + {}_2r_1$$

or

(9) $${}_2\rho_1 > {}_2r_1$$

Inequality (9) implies that the trader should now make one-period forward loans for two periods hence. Again, in the perfect world of this example, the transaction can be accomplished by selling short two-period bonds and using the proceeds to buy three-period bonds. *After two periods*, this will bring an expected capital gain of

(10) $$d_2\left[\frac{(1 + R_3)^3}{(1 + {}_2r_1)} - (1 + R_2)^2\right]$$

where d_2 is the dollar amount of both the long and short transaction. By substituting for $1+R_2$ from (H.1) and using inequality (H.4), one can verify that (10) is also positive.

In summary, the present decision rule instructs a trader to make the following transactions at period zero:

(a) Buy one-period bonds with available resources.
(b) Sell short three-period bonds and use the proceeds to buy one-period bonds.
(c) Sell short two-period bonds and use the proceeds to buy three-period bonds.

Transaction (b) is kept open for one period, period 0 to 1, and transaction (c) is kept open

for two periods. Only transaction (a) was recommended by Wood's decision rule (2) and none of the three transactions were recommended by (1). The positive expected capital gains of (8) and (10) that accrue to a trader using the revised decision rule prove that rules (1) and (2) are sub-optimal.

The quantities of bonds bought and sold in transactions (b) and (c) are unspecified. This is a very important fact that requires elaboration. If the trader is truly risk indifferent, and really wants to maximize *expected* return, he would attempt to make d_1, in expression (8) and d_2 in (10) as large as possible. Only by transacting an infinite quantity of bonds would he maximize expected return. The fact that we rarely observe investors attempting to trade infinite amounts brings out the unrealism of the preceding example. Markets are not perfect in the special sense used there and traders can neither grant nor issue unlimited quantities of loans. Even if they could, it is likely that none *would* because no trader operating with unlimited liability is completely indifferent to the risk of total ruin.

The example is important, however, in clarifying the central point of this comment: Expected future spot rates do have behavioral significance. By comparing them to current forward rates, the bond trader chooses an optimal investment strategy. With the introduction of uncertainty, their comparison to forward rates acquires an even more crucial role. Assuming no risk of default, the forward rate is perfectly certain whereas the corresponding future spot rate is a random variable.[6]

REFERENCES

I. Fisher, *The Nature of Capital and Income*, London 1912.

J. R. Hicks, *Value and capital*, 2d ed., Oxford 1946.

F. A. Lutz, "The Structure of Interest Rates," *Quart. J. Econ.*, Nov. 1940, *40*, 36–63, rep. in AEA, *Readings in the Theory of Income Distribution*, Homewood 1951, 499–529.

D. Meiselman, *The Term Structure of Interest Rates*, Englewood Cliffs 1962.

R. Roll, "The Efficient Market Model Applied to U. S. Treasury Bill Rates," unpublished doctoral dissertation, Univ. Chicago 1968, forthcoming as *The Behavior of Interest Rates: An Application of the Efficient Market Model to U.S. Treasury Bill Rates*, New York 1970.

J. Wood, "Expectations and the Demand for Bonds," *Amer. Econ. Rev.*, Sept. 1969, *59*, 522–30.

[6] The implications of this environment have been worked out in my doctoral thesis.

Expectations and the Demand for Bonds: Comment

By A. BUSE*

The traditional theory of the term structure of interest rates, as exemplified by J. R. Hicks, is usually considered to be a coherent and consistent doctrine. It is not the logic of the theory that has been questioned by the critics but its empirical relevance; expectations to Kingdom Come being the more humorous manifestation of this attitude. However, in a recent stimulating paper in this *Review*,[1] John Wood has asserted that the decision rule implied by the traditional theory leads to sub-optimal results (as measured by holding period returns) unless the investor is assumed to hold each security until maturity. Without this assumption, Wood argues, the optimal decision rule re quires that investors make forecasts of interest rates for only one period into the future, unlike the traditional theory which involves forecasts of expected one-period rates for n periods into the future. A three-period example is used to demonstrate the result.

The purpose of this comment is to demonstrate that Wood's sub-optimality result does not hold. Although the algebra of his three-period example is correct, the result he obtains depends on the unstated assumption that the Wood-type investor makes correct forecasts and the traditionalist does not. This being the case, it is hardly surprising that the traditionalist is sub-optimal. If the forecasting abilities are reversed, so are the optimality relations.

Wood's three-period example, (pp. 526–28), can be used to demonstrate the assertions made in the previous paragraph. The point of departure is a pair of inequalities (H.4) and (G.4) which specify the divergence between existing market rates and those of the H-investor (traditional decision rule) and the G-investor (Wood decision rule).

(H.4) $(1+{}_0R_3)>[(1+{}_0R_1)(1+{}_1r_1)(1+{}_2r_1)]^{1/3}$

(G.4) $(1+{}_0R_3)<(1+{}_0R_1)^{1/3}(1+{}_1r_2)^{2/3}$

The current one- and three-period market rates[2] are ${}_0R_1$ and ${}_0R_3$. As of period zero, ${}_1r_1$ and ${}_2r_1$ are the one-period rates investor H expects in periods one and two. It follows that the H-investor expects the two-period rate in period one to be

$$[(1+{}_1r_1)(1+{}_2r_1)]^{1/2}-1$$

The G-investor, who makes forecasts of rates for only one period into the future, expects the two-period rate to be ${}_1r_2$ in period one. G and H are assumed to have identical expectations about the one-period rate in the next period.

Given these conditions, investor H will buy three-period bonds and G will buy one-period bonds. Wood then compares the returns obtained by H and G after the lapse of one period. Obviously, the return to G is ${}_0R_1$ per unit of investment. The return to H will depend on the price of the three-period security at the start of period one. If this price is denoted by ${}_1P_2$, the one-period return to H is given by

$$({}_0P_3/{}_1P_2)-1=[(1+{}_0R_3)^3/(1+{}_1R_2)^2]-1$$

Wood (p. 527) states that the return to H is given by

$$(1+{}_0R_3)^3/(1+{}_1r_2)^2-1,$$

which is less than ${}_0R_1$ by (G.4). Wood's result follows if ${}_1r_2={}_1R_2$; that is, G's expectation of the two-period rate for period one is the rate that actually prevails in the market

* Associate professor of economics, University of Alberta.

[1] Although I try to demonstrate that Wood's results are in error, I found the paper stimulating nonetheless because I was forced to think carefully about the implications of the traditional expectations theory.

[2] A minor change in notation from the original has been made in order to make the argument as explicit as possible. Thus, ${}_tR_j$ is the observed yield on a j-period bond in period t. Similarly, ${}_tP_j$ stands for the observed price of a j-period bond in period t.

at t = 1. In short, G's expectations are correct. If this is the case, then H's expectation of the two-period rate must be wrong since by (H.4) and (G.4). (Wood's inequality (7))

$$(1)\qquad {}_1r_2 > [(1 + {}_1r_1)(1 + {}_2r_1)]^{1/2} - 1$$

where the right-hand side is H's expectation of the two-period rate.

Assume now that H's expectations are correct so that after a lapse of one period $(1+{}_1R_2)^2 = (1+{}_1r_1)(1+{}_2r_1)$. Then the holding period return to H is given by

$$(2)\qquad [(1 + {}_0R_3)^3/(1 + {}_1r_1)(1 + {}_2r_1)] - 1$$

which by (H.4) is greater than ${}_0R_1$. The G-investor now has the lower rate of return and his decision rule appears to be sub-optimal.

Another possible way of recognizing Wood's implicit assumption is to consider his description (p. 527) of the inequality (7)

$$(3)\qquad (1 + {}_1r_2)^2 > (1 + {}_1r_1)(1 + {}_2r_1)$$

which is implied by (H.4) and (G.4). He states that "Inequality (7) shows where H's expectations differ from the forward rates implicit in current rates." Clearly he is treating ${}_1r_2$ as the implied market forward rate (to be subsequently realized) when in point of fact the current rates imply a two-period rate in period one equal to[3]

$$[(1 + {}_0R_3)^3/(1 + {}_0R_1)]^{1/2} - 1$$

[3] It is of course possible that the "market" has correct expectations so that $(1+{}_0R_3)^3/(1+{}_0R_1) = (1+{}_1R_2)^2$. In this case the realized returns to H and G are equal although H's realized return is less than his expected return. G could be similarly disappointed if the inequalities in (H.4) and (G.4) were reversed.

which by (G.4) is not equal to ${}_1r_2$. It can also be noted that inequality (7) is inconsistent with Wood's assumption that H and G have identical expectations but this inconsistency is secondary to the main point that the G decision rule does not possess the dominance property attributed to it by Wood.

To sum up, Wood has not proved that a decision rule based on forecasts of the long rates in the next period is superior to a decision rule which forecasts the expected one-period rates. It follows that he has not demonstrated that the traditional expectations theory is incomplete without the assumption of bond holding to maturity. These conclusions are not that surprising since Wood indicates that the decision rules lead to different actions only when inequalities such as (7) hold. Such inequalities specify differences in forecasts of future rates and the results of any action will depend on the success of the forecast. The higher return will always accrue to the better forecaster. This is not to say, however, that the decision rule proposed by Wood does not have substantial intuitive appeal. Such a rule may in fact be relevant to a successful explanation of the behavior of the term structure. But this is an empirical issue, not a logical one. Whether the issue will be resolved next year or by Kingdom Come is another matter.

REFERENCES

J. R. Hicks, *Value and Capital*, 2d ed., Oxford 1946.

J. H. Wood, "Expectations and the Demand for Bonds," *Amer. Econ. Rev.*, Sept. 1969, *59*, 522–30.

Expectations and the Demand for Bonds: Comment

By Reuben A. Kessel*

The thesis of John Wood's article is: Predicting the next period price of a bond constitutes a more correct criterion, i.e., leads to better results for the investor than predicting the expected spot rates over the life of a bond which is the classical view. It is the purpose of this note to argue that this constitutes a distinction without a difference and the dissimilar results Wood obtains in comparing the two criteria is a consequence of his failure to understand the process of arbitrage and speculation in the world he postulated, a world of zero transactions costs.

Wood's proof that his criterion is better than the criterion he imputes to Fisher, Lutz, and Hicks rests on the following example: If the average of the expected one-period rates one and two periods hence is less than the expected two-period rate one period hence, then directly estimating the price of a bond one period hence will yield a different and better guide to action than estimating the expected one-period rates into the indefinite future.

Symbolically,[1]

(1) $(1 + {}_1r_2)^2 > (1 + {}_1r_1)(1 + {}_2r_1)$

In addition,

(2) $(1+R_3) > [(1+R_1)(1+{}_1r_1)(1+{}_2r_1)]^{1/3}$

and

(3) $(1 + R_3) < (1 + R_1)^{1/3}(1 + {}_1r_2)^{2/3}$

According to Wood, Hicks et al. will choose the three-period rate and earn $\frac{(1 + R_3)^3}{(1 + {}_1r)^2}$ in the first period, whereas Wood would buy one-period bonds and earn $(1+R_1)$ which, from (3), is greater than $\frac{(1 + R_3)^3}{(1 + {}_1r_2)^2}$.

It is at this point that Wood commits a crucial error. In a market in which there are zero transactions costs, the inequality between the geometric mean of expected one-period rates one and two periods hence and the expected two-period rate one period hence cannot hold; this market is not in equilibrium. Under the conditions postulated, arbitrage will occur. Arbitrageurs will buy forward two-period money for delivery one period hence and sell forward for delivery one and two periods hence one-period money. As a result, the inequality will be eliminated. This arbitrage opportunity holds in a world of uncertainty as well as perfect forecasting. Once the inequality is eliminated, both criteria will yield identical answers.[2]

It should come as no surprise that both criteria yield identical answers. The processes by which investment decisions are made are identical. To obtain the expected price one period hence, one must look at the forward rates over the life of the securities outstanding. This is the same process that Lutz, Hicks, and Fisher postulate.[3]

* Professor of business economics, Graduate School of Business, University of Chicago.

[1] The expected rate is r. The prescript represents the number of periods that must elaspse before a rate becomes a spot rate; the postscript represents the number of periods a rate encompasses. R depicts spot rates.

[2] This same point is relevant for understanding the implications of the other inequalities Wood considered. Forward rates undervalued in the market will be bought and conversely.

[3] There are important differences between Hicks and Lutz that Wood ignores. Uncertainty is integral to the views of Hicks; liquidity preference and certainty cannot logically coexist.

Expectations and the Demand for Bonds: Reply

By JOHN H. WOOD*

The first half of Richard Roll's comment contains an argument that my interpretation of the traditional expectations hypothesis of the term structure of interest rates is incorrect. The second half argues that both the decision rule that I infer from the traditional theory and the decision rule that I have advanced as an improvement upon that approach are sub-optimal. Adolf Buse and Reuben A. Kessel, on the other hand, accept my interpretation of the received theory and argue that the traditional approach is correct. I will respond first to Roll's criticism of my interpretation (and that of Buse and Kessel) of the traditional literature.

I. Fisher, Hicks, and Lutz

Any dispute concerning Irving Fisher's understanding of the decision process by which investors choose among bonds of different maturities ought to be settled by quoting his complete statement, of which Roll gives us the final sentence. The passage immediately preceding the sentence quoted by Roll is as follows:

> If the intention in advance is to reinvest, it becomes important not simply to know the present rate of interest, but to forecast the future rate. This enters into the calculations of an investor who holds a 25-year bond at 5 percent. He will usually regard the final payment as 'principal', intending that when it becomes due it shall be reinvested in a similar 25-year bond. He, therefore, is not really buying a 25-year income stream of $5 a year plus $100 at the end of the term, but is buying let us say, a *50*-year income stream consisting of $5 per year for the first 25 years and an *unknown* amount per year during the second 25 years. In order to forecast what income will be received in the second period, he has to forecast the rate of interest. In other words, although the bond represents nominally a fixed and certain series of income items, yet, in view of the intention to reinvest, it actually represents an income which is quite uncertain after 25 years, because of the uncertainty in the future rate of interest. Such an investor, if he expected the rate of interest at the end of 25 years to be 2 percent, would, in purchasing the above-mentioned bond, be getting $5 a year for 25 years and $2 a year for the next 25 years. Under these conditions, if he could buy a 50-year bond at 4 percent, he would prefer to do so. But, if he expected the rate of interest to remain, for each 25-year period, at 5 percent, he would prefer, rather than invest now in a 50-year bond at 4 percent, to invest in the 25-year bond at 5 percent, intending to reinvest at 5 percent at the expiration of the term. His forecast of what the rate of interest will be in 25 years will thus materially affect the choice of investments to-day.[1] [1906, pp. 273–74]

Fisher's discussion clearly indicates that in choosing between current purchases of long- and short-term bonds the investor forecasts interest rates many years in the future. He then compares the current rate

* Finance department, University of Pennsylvania. I wish to take the opportunity in this reply to remedy an omission from my article. Herschel Grossman has published a paper in which he also develops at some length the point that, under conditions of certainty and zero transactions costs, an investor who desires to maximize wealth over any horizon utilizes forecasts of interest rates only one period in the future. Unfortunately, Grossman's contribution came to my attention too late for me to refer to it in my article.

[1] This quotation also makes clear that, when Fisher writes about expectations of rises and falls in the rate of interest, he is referring to the rate of interest on 25-year bonds that is expected to prevail 25 years in the future. In Fisher's example, this is the short-term rate. Hence, Roll is incorrect when he inserts "long-term" in brackets to modify "rate of interest" in the first sentence of his quotation.

on long-term bonds with the average of current and future rates on short-term bonds. If the former exceeds the latter, he buys long-term bonds; if the relationship is reversed, he buys short-term bonds; if the equality holds, he is indifferent between long- and short-term bonds.[2] Such an action based upon such a comparison is the essence of a decision rule. So far as I am aware, this passage from Fisher is by far the clearest and most complete description of the decision process underlying the equilibrium relationship among interest rates implied by the expectations hypothesis. The quality of discussions of that theory, even by Fisher,[3] deteriorates rapidly after 1906.[4] For example, after stating his assumptions of certainty, no transactions costs, and complete shiftability among bonds by both borrowers and lenders, Lutz gives the equilibrium relationship among rates that he asserts to be implied by these assumptions in a footnote with no intervening discussion (pp. 499–500). The equation given by Lutz (who credits Fisher, Hicks, and others with the development of the theory) is simply an algebraic expression of Fisher's argument for the special case in which the equality holds. The remainder of Lutz's Section I, which is devoted to the basic theory (i.e., where certainty and zero transactions costs prevail), consists solely of a series of algebraic manipulations of the formula contained in his footnote with the added implicit assumption of zero elasticity of expectations. Roll's quotation from the Lutz article, which is taken from Section IV (pp. 512–20) is not relevant to the criticism presented in my paper, which is directed toward the Fisherian theory that Lutz accepted without question in his Section I (pp. 499–504). The discussion in Section IV of Lutz presents an analysis of the determination of the equilibrium structure of rates in a world in which Fisher's theory has been modified by the introduction of transactions costs (in Section II) and uncertainty (in Section III).[5] Consequently, the passage quoted by Roll is taken from the discussion of a theory very different from that of Fisher's and very different from that which has been accepted as "the" expectations hypothesis, the cornerstones of which are the assumptions of certainty and zero transactions costs. It will further become clear to the reader of Lutz's discussion in its entirety that one of the similarities between Sections I and IV is that the investor is required to forecast short rates several periods in the future, though not necessarily in the analysis of Section IV over the entire life of the long-term investment. Contrary to the impression given by Roll's discussion and his quotations from Fisher and Lutz, both of these writers require investors to forecast rates several periods into the future.

Roll's assertion with respect to Hicks' treatment of the term structure of interest rates is also incorrect. As the reader may verify, Hicks (pp. 144–45) begins with a statement of the equilibrium relationship among rates in the absence of risk. This relationship is identical to that implied by Fisher's discussion ("if no interest is to be paid until the conclusion of the whole transaction" (Hicks, p. 145)) and is expressed in its general form in equation (1) of my article. It is no accident that the terminology preceding my equation (1) is very similar to that used by Hicks. Only after his statement without qualification of the theory received

[2] Compare Fisher's discussion with my statement of the traditional theory (pp. 522–24) which was drawn directly and without modification from the passage just quoted and from the equivalent (when the equality holds) passage in J. R. Hicks (pp. 144–45).

[3] See (1930, p. 70) in which Fisher merely asserts without defense and without reference that "a rate on a five year contract may be considered as a sort of an average of five theoretically existing rates, one for each of the five years covered."

[4] The most notable exception is Burton Malkiel (pp. 18–20), who discusses the decision process implied by the traditional theory in detail. J. W. Conard (p. 294) also alludes to a decision process of the type attributed in my article (pp. 523–24) to that theory.

[5] The precise manner in which uncertainty and transactions costs enter Lutz's modification of the traditional theory is not, however, at all clear. Lutz is extremely vague about the assumptions underlying the analysis of his Section IV. See D. G. Luckett for a valiant attempt to decipher this portion of Lutz's article.

from Fisher does Hicks (like Lutz) introduce uncertainty into the analysis.[6]

II. Roll's Decision Rule

Roll refers to the three-bond example of my article, where I showed that an investor using the decision rule of the traditional theory "prefers three-period over one-period securities," whereas an investor following the decision rule that I advance, which utilizes forecasts only one period in the future, "prefers one-period over three-period securities" (p. 527). Roll agrees with my point that the investor should "(a) make one-period spot loans now with available resources." But he goes further than I did in his recommendation (b) that the wealth maximizing investor should also issue three-period loans now in order to buy one-period bonds with the proceeds. I agree. This is a useful extension of my results. Actions (a) and (b) are both desirable if Roll's expression (8) is positive, which follows from the inequality (G.4) on page 527 of my article. (Also see the inequality on page 528.) As I emphasized, (G.4) utilizes forecasts only one period in the future.

Roll's recommendation (c), however, is sub-optimal because at this point he commits the classic Fisher-Hicks-Lutz error. He uses the Hicksian decision rule (H.4) to derive the result that the investor should issue two-period bonds and use the proceeds to buy three-period bonds. "*After two periods*, this will bring an expected capital gain" equal to the amount shown in (10). But Roll has already shown in (a) and (b) that one-period bonds are currently preferable to three-period bonds. Why should an investor issue three-period bonds in order to raise funds with which to buy one-period bonds, as in (b), and then in (c) issue two-period bonds in order to buy three-period bonds?[7] Action (c) is consistent with wealth maximization only if the investor is bound to hold securities for two-periods. Otherwise, the investor will never buy three-period bonds currently regardless of the source of funds. Roll seems to be on the verge of an explicit recognition of this condition when he says that action (c) brings an expected capital gain "*after two periods*." But if we constrain investors to hold securities until maturity so that (c) is optimal, then (a) and (b) cease to be optimal. This is precisely the point made in my article.[8]

Roll's point in his next-to-last paragraph is valid and is one that I recognized in my article.[9] The infinite elasticity of demand functions that follows from the assumptions of certainty, zero transactions costs, and perfect competition is one of the more troublesome aspects both of the traditional theory and of the theory advanced in my article. In the case of divergent expectations, market equilibrium solutions are rendered indeterminate unless "We assume . . . that the command over resources by any individual investor is limited . . . " (p. 523, fn. 5). On the other hand, as Meiselman (pp. 10, 54, 57) has argued, perfectly elastic demand functions may actually strengthen the theory if expectations are identical, which is an assumption usually associated with the expectations hypothesis (see Lutz, p. 499, Conard, p. 290, Malkiel, p. 18, and Wood, p.

[6] Hicks' attempt to deal with the term structure under uncertainty, like that of Lutz, was unsuccessful. Hicks imposed "liquidity premiums" on a theory which has since been shown to be invalid (see Wood) and with no discussion of the objective functions to be maximized by investors or of the nature of the constraints confronting them. With one exception, those studies of the term structure under uncertainty that explicitly use techniques of constrained optimization have been limited in effect to comparisons between one- and two-period bonds (see G. O. Bierwag and M. A. Grove, and H. A. J. Green), which is the only case in which the traditional theory is valid (see Wood p. 526) and, hence, the only case which can support the kinds of rigorous analysis under uncertainty such as have been pursued by Bierwag and Grove and by Green. The exception is Grossman who is able to deal with bonds of varying maturity because he recognizes that the traditional theory is invalid and discards it from the beginning of his analysis.

[7] Note that in (a) and (b), Roll prefers one-period to three-period bonds and then in (c) prefers three-period to two-period bonds when it is clear from (H.1) that by both the Hicksian and one-period forecasting rules, the investor is indifferent between one- and two-period bonds. He gets these non-transitive results because he jumps back and forth between contradictory decision rules.

[8] See the last paragraph of Section II, p. 528.

[9] See fn. 5, p. 523. Also see Don Patinkin p. 68.

522). But Roll's suggestion that our understanding of the structure of interest rates will be improved by the introduction of uncertainty into our analysis is good advice. On this point, see footnote 6 above and the last paragraph of my article.

III. Buse: Contradictory Expectations?

Buse argues that in my three-bond example investors pursue different courses of action not because of different decision rules but because of differences in forecasting. It is difficult to respond to this criticism except by pointing out that I explicitly assumed identical expectations (p. 527). Further, it is easily seen that the substitution of some configuration of expected rates into the decision rule (2) that Buse agrees is implied by the traditional theory (p. 524) may yield a different result from that obtained by substituting the same expectations into the decision rule (5) that utilizes forecasts only one period in the future (p. 525). My three-bond example is merely "the simplest case in which decision rules (2) and (5) produce different results . . . " (p. 526). Can anyone, Buse excepted, doubt that different decision rules may yield different results even though the same information is available to both?

Buse obtains his results that investor *G* sometimes does better and sometimes worse than *H* by indiscriminately jumping between the assumption that (1) is an inequality and the assumption that (1) is an equation. Confusion results especially from his substitution of the equation form of (1) into the inequality (H.4) and then comparing the result with the inequality (G.4). Buse performs this comparison despite my demonstration (p. 526) that, given identical expectations and indifference between one- and two-period bonds by *H* and *G* (assumptions explicitly underlying both my equilibrium and disequilibrium examples), if (1) is an equation then (H.4) and (G.4) reduce to equations (H.2) and (G.2) which are equivalent. Naturally, Buse's different and sometimes contradictory assumptions yield contradictory results. I retained the same set of assumptions throughout the course of my discussion of the disequilibrium case (including the assumption that the inequality (1) holds) and hence obtained a unique result. Buse's mistake is the common one in discussions of the expectations hypothesis of confusing equilibrium and disequilibrium situations.

IV. Kessel: Identical Result in Equilibrium

Buse's last line of defense consists of the same argument as that advanced by Kessel. That argument is that the decision rule implicit in the traditional theory and the one suggested in my article yield identical results in equilibrium, i.e., when the inequalities in my example become equations. Kessel's argument consists of the following four statements: (i) The example discussed in my paper describes a position of market disequilibrium; (ii) Forces will be set in motion that will restore equilibrium, in which case (1)–(2) become equations; (iii) In equilibrium, both the Hicksian and one-period forecasting rules "yield identical answers"; (iv) Consequently, "to obtain the expected price one period hence, one must look at the forward rates over the life of the securities outstanding." Kessel and I are in complete agreement with respect to the first three statements.[10] But the fourth statement is a non sequitur. One wonders why, even in equilibrium, if the two criteria yield identical answers it is necessary to forecast rates many periods into the future when by Kessel's admission the one-period forecasting model does as well. In fact, any decision criterion, including a dart board, yields an optimal result in equilibrium.

Neither Kessel nor Buse perceives the nature of the distinction from the standpoint of decision makers between equilibrium and

[10] It was shown in my article that, during the period of adjustment, an investor using the one-period forecasting rule will earn at least as much and sometimes more than an investor using the traditional rule, a result not disputed by Kessel. For discussions of movements from disequilibrium to equilibrium positions that are consistent with Kessel's description of the adjustment process, see pages 523–24, 528–29 of my article. Also see Malkiel pp. 19–20. Kessel's statement (iii), which has served for many years as the clinching argument in support of the traditional theory, is discussed in my article on pages 524, 526–27.

disequilibrium situations. Under the assumptions stated at the beginning of my article, there is no investment problem when the market is in equilibrium (i.e., when inequalities such as (1)–(2) are eliminated) and hence no need for an investment criterion because all investments are expected to yield identical one-period returns. In a world in which all markets were somehow continuously in equilibrium, neither the Hicksian nor the one-period forecasting rule would be of any use; neither would have any behavioral significance and the forward rates implicit in both would be mere tautologies. Consequently, the two rules are meaningful only in regimes in which disequilibria are possible. And it was shown in my article that, under the assumptions generally thought to underlie the traditional theory, the one-period forecasting rule dominates the rule implicit in the traditional theory in disequilibrium situations. Thus, under conditions of certainty and zero transactions costs, wealth maximizing investors make decisions on the basis of forecasts only one period in the future. The result is that the forward rates more than one period in the future defined by the observed structure of rates are without behavioral significance.

An analogy drawn from the theory of the firm will illustrate my point regarding the importance of distinguishing between market equilibria and disequilibria in the development of decision rules. In long-run competitive equilibrium where entry is free, profit maximizing firms will produce quantities such that marginal and average costs are equal to each other and to the price of the product. If the industry were continuously in long-run equilibrium, the following decision rules would yield identical and optimal results: (a) operate such that marginal cost equals price (subject to the condition that marginal cost is rising); (b) minimize average cost. But to say that (a) and (b) yield identical results in equilibrium does not obscure the fact that in disequilibrium situations (i.e., nearly all of the time), (a) dominates (b).[11]

REFERENCES

G. O. Bierwag and M. A. Grove, "A Model of the Term Structure of Interest Rates," *Rev. Econ. Statist.*, Feb. 1967, *49*, 50–62.

J. W. Conard, *An Introduction to the Theory of Interest*, Berkeley 1959.

I. Fisher, *The Nature of Capital and Income*, New York 1906.

———, *The Theory of Interest*, New York 1930.

H. A. J. Green, "Uncertainty and the Expectations Hypothesis," *Rev. Econ. Stud.*, Oct. 1967, *34*, 387–98.

H. I. Grossman, "Expectations, Transactions Costs, and Asset Demands," *J. Finance*, June 1969, 491–506.

J. R. Hicks, *Value and Capital*, 2d ed., Oxford 1946.

R. A. Kessel, "The Cyclical Behavior of the Term Structure of Interest Rates," Nat. Bur. Econ. Res. Occas. Paper 91, New York 1965.

D. G. Luckett, "Professor Lutz and the Structure of Interest Rates," *Quart. J. Econ.*, Feb. 1959, *73*, 131–44.

F. A. Lutz, "The Structure of Interest Rates," *Quart. J. Econ.*, Nov. 1940, *55*, 36–63; rep. in AEA *Readings in the Theory of Income Distribution*, Homewood 1946, 499–529.

B. G. Malkiel, *The Term Structure of Interest Rates*, Princeton 1966.

D. Meiselman, *The Term Structure of Interest Rates*, Englewood Cliffs 1962.

D. Patinkin, *Money, Interest and Prices*, 2d ed., New York 1965.

J. H. Wood, "Expectations and the Demand for Bonds," *Amer. Econ. Rev.*, Sept. 1969, *59*, 522–30.

[11] With respect to Kessel's fn. 3, see my response above to Roll's criticism of my reference to Hicks.

Output of the Restrained Firm: Comment

By A. Ross Shepherd*

In a recent article in this *Review*, Milton Z. Kafoglis examined the price and output behavior of the restrained monopoly firm and found that when firms seek to maximize output or scale of operations, "... the output of the restrained monopoly may exceed that predicted by existing models and may even be pushed beyond optimum (in the Paretian sense) as a result of sales at prices below marginal cost" (p. 583). *Inter alia*, Kafoglis finds that even in single markets under increasing cost Pareto optimal output "... will be exceeded by the output-maximizing firm and if demand elasticity exceeds unity in the range of restraint, by the revenue-maximizing firm" (p. 586). The purposes of this comment are: (1) to reveal the implicit assumption on which Kafoglis' result is based; (2) to show for the increasing cost case the assumption under which constrained output and revenue maximizers will produce the optimal output; and (3) to show that even in single markets *monopsony* power will likely result in sub-optimally *large* outputs for constrained output and revenue maximizers.

Figure 1 shows the monopolist's average revenue (*AR*) and marginal revenue (*MR*) curves together with various long-run cost curves emanating from point *M*.[1] Compared to Kafoglis' Figure 1 (p. 584), curve *MA* corresponds to the rising portion of Kafoglis' average cost curve (*AC*) and *MB* gives marginal cost (*MC*) above average cost. Thus in

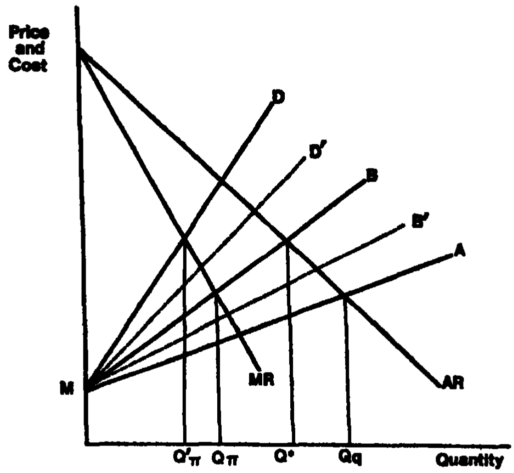


FIGURE 1

Figure 1, *MA* is the long-run average cost curve and *MB* is the curve marginal to *MA*. According to Kafoglis, Q_π is the profit-maximizing output, Q^* is the Pareto optimal output, and Qq is the maximum output consistent with a break-even profit constraint. Now Q^* is Pareto optimal only if *MB* gives the cost to *society* of producing and selling the marginal unit. In order for Q_π to be the profit-maximizing output, *MB* must also give the increment to the *firm's* total cost that results from the production and sale of the Qth unit. These two conditions are consistent if, and only if, the firm pays no rent to factors of production. This is possible if factor prices are invariant with respect to the level of the firm's activity; it is also possible if the firm is a perfectly discriminating monopsony. From the general context of Kafoglis' discussion it would appear that he implicitly assumes fixed input prices.

It seems clear that an adequate analysis of industry behavior must include the possibility that industry expansion occurs under conditions of unconstrained substitution among inputs and rising factor supply

* Associate professor of economics, University of Missouri-Kansas City. I wish to thank the managing editor for helpful comments.

[1] As Dean A. Worcester, Jr. has pointed out (fn. 9, pp. 879–80), the convention of drawing U-shaped cost curves for the monopoly firm can be misleading. The monopoly firm is the analytical counterpart of the competitive industry, and the long-run cost curves in each case will likely come from the same family. For some purposes the conventional U-shaped curve is convenient in that it illustrates all the long-run cost possibilities. However, the present discussion is limited to the increasing cost case, so we show only rising cost curves.

prices. When rising average costs reflect rising input prices the monopolist is altogether likely to recognize that he has monopsony power, which means that he will see that his marginal costs include increases in intramarginal factor rents. From society's standpoint, however, these intramarginal rent increases are not costs; only the cost requisite to induce the marginal factors to produce and market the marginal unit of output is a cost to society. The implications of rising factor supply prices for the behavior of the restrained monopoly firm are clearly seen after an appropriate reinterpretation of Figure 1.

In the context of rising factor prices, curve MA is average cost exclusive of factor rents; MB, which is mathematically marginal to MA, is both marginal cost exclusive of factor rents, and average cost including average competitive factor rents. Under perfect competition MB is the long-run average cost or supply curve of the industry, and MA is analytically irrelevant because no competitive firm can avoid paying factor rents. In the presence of monopsony power, however, MA is to be interpreted as the average cost curve of the perfectly discriminating monopsonist, for whom MB is the curve of the marginal costs. Curve MD is marginal to MB and gives under perfect competition the total increase in cost, including the increase in intramarginal factor rents, associated with the production and sale of the Qth unit.[3]

Consider now the familiar, illustrative case of a product produced with two factors. It is well known that a least-cost input combination requires:

$$\text{(1)} \qquad \frac{MPP_1}{MFC_1} = \frac{MPP_2}{MFC_2},$$

where MPP denotes marginal physical product, MFC is marginal factor cost and the subscripts indicate the respective factors. The relationship between MFC, factor price (P) and the factor supply elasticity (E) is also well known:

$$\text{(2)} \quad MFC_i = P_i(1+1/E_i) \; (i=1, 2)(E\neq 0)$$

In the case of perfect competition, each firm's perspective is such that for it $E_i = \infty$ and $MFC_i = P_i$, and the invisible hand will lead the competitive industry to use inputs such that $MPP_1/P_1 = MPP_2/P_2$. Now if supply elasticities are not infinite but are equal, ratios of MFC's still equal ratios of P's and least-cost combinations for given outputs are not altered by perceived monopsony power. In this case the monopoly-monopsonist's cost curves will coincide with those of the competitive industry (provided that monopolization of a competitive industry does not yield other net private economies or diseconomies). In Figure 1 the profit-maximizing monopoly-monopsony will aim at output Q while the constrained output and revenue maximizers will produce the optimal output, Q^*. We have it, then, that when rising output costs reflect increasing input prices, the constrained output and revenue maximizers will produce a socially optimal rate of output if factor supply elasticities are equal. In what follows it will be shown that for this optimal result to obtain equal factor supply elasticities are necessary as well as sufficient.

Factor supply elasticities will usually differ and, as compared to the competitive industry, the monopoly firm with monopsony power (hereafter, simply "monopoly") will economize on the relatively inelastic factor. This will both reduce the monopolist's (private) costs and misallocate resources. For some rate of output Q let:

[3] It is worthwhile to specify the assumptions that underlie the interpretations of this paragraph. 1) Factors are homogeneous as viewed by the industry for whom the cost curves are drawn, but heterogeneous from the standpoint of alternative uses. This ensures rising marginal transfer prices and the payment of factor rents to intramarginal units. 2) Expansion of the industry does not increase the prices of the factors in alternative uses so that the transfer cost of any given unit of a factor is independent of the level of the industry's output. 3) The industry production function exhibits constant technical returns to scale, and there are no pecuniary economies of large-scale production. These latter two assumptions ensure coincidence between the curves showing marginal competitive cost exclusive of factor rents and average competitive cost including average factor rents. The *loci classici* for all this are Joan Robinson's chapters 8 and 10.

$$(3) \qquad \frac{MPP_1}{P_1(1+1/E_1)} \neq \frac{MPP_2}{P_2(1+1/E_2)},$$

where

$$(4) \qquad \frac{MPP_1}{P_1} = \frac{MPP_2}{P_2}$$

A competitive least-cost combination obtains, and the output is produced with the socially optimal combination of inputs because the competitive industry has ignored intramarginal factor rents, which are not social costs. However, the monopolist will note that he can reduce the private cost of this output by economizing on the less elastic factor until extra output per last dollar spent is the same for each input. Hence, if at every point on the competitive supply curve factor supply elasticities differ, the monopolist's average cost will be below the competitive level for all outputs greater than zero.

In Figure 1 MB' and MD' are, respectively, the monopolist's average and marginal cost curves when factor supply elasticities differ. It is clear that monopsonistic private cost saving will lead the constrained output and revenue maximizers beyond Q^*, the social optimum, while the profit maximizer will move closer to Q^*.[3] The ultimate in monopsonistic cost saving is realized by the perfectly discriminating monopsonist, for whom the marginal cost is MB and average cost is MA. This brings us back to the zero rent results of Kafoglis.[4]

REFERENCES

M. Z. Kafoglis, "Output of the Restrained Firm," *Amer. Econ. Rev.*, Sept. 1969, *59*, 583–89.

J. Robinson, *The Economics of Imperfect Competition*, London 1933.

D. A. Worcester, Jr., "Pecuniary and Technological Externality, Factor Rents, and Social Cost," *Amer. Econ. Rev.*, Dec. 1969, *59*, 873–85.

[3] In the case of the profit maximizer the output adjustment resulting from this monopsonistic misallocation of resources counteracts in some measure the social distortion of monopsonistically contrived scarcity.

[4] It will be noted that if the firm is both a perfectly discriminating monopsony and a perfectly discriminating monopoly, the profit maximizer will produce a socially optimal rate of output.

Output of the Restrained Firm: Reply

By Milton Z. Kafoglis*

A. Ross Shepherd correctly reasons that the monopoly firm examined in my original analysis must expand with fixed input prices and that the average cost curve must, therefore, exclude rent to fixed factors. In relaxing this assumption to allow for rising input price (and the payment of full rent), Shepherd introduces the possibility that the output (or revenue) maximizing firm may produce the Pareto optimal output. We should conclude therefore that in the case of single markets under increasing cost the output maximizing monopoly will attain Pareto optimal output if the average cost curve coincides with the competitive supply schedule but will exceed Pareto optimal output if the average cost curve lies below the competitive supply schedule. In the ordinary case we would predict a lower average cost curve because the firm probably can avoid the payment of full rent and may practice various forms of discrimination in factor markets.

It would seem that the possibility of rent avoidance is especially significant in the case of regulated industries where effective calculations of opportunity costs are frustrated by legal and regulatory barriers which prevent consideration of alternative uses and by rate base calculations in terms of historical money costs. In the case of publicly-owned utilities this situation may be aggravated by tax exemptions and other concessions. Supported by such institutions the firm may be expected to develop a cost schedule which fails to reflect the full opportunity costs of the resources employed and, if restrained only by the fair return criterion, will be able to expand output beyond Pareto optimal.

Shepherd's comment is especially instructive because it brings attention to the fact that the level of the average cost curve is affected by industrial structure. However, since most of these effects will be intramarginal, the marginal cost curve may not be altered significantly and analysis on the traditional assumption of profit-maximization may yield correct price and output predictions. On the other hand, output and price may be affected significantly by the level of the average cost curve in the case of restrained firms. In such cases it becomes necessary to be more explicit about intramarginal factor payments than my original treatment.

* Professor of economics, University of Florida.

Production Indeterminacy with Three Goods and Two Factors: A Comment on the Pattern of Trade

By DOUGLAS B. STEWART*

James Melvin's examination of the indeterminacy in the three-good, two-factor, two-country trade model prompts him to claim in his recent article in this *Review* that whenever all goods are traded, that country exporting the labor intensive good will also be exporting the capital intensive good (p. 1263). Recognizing the damage this claim does to the standard Heckscher-Ohlin theorem, Melvin reformulates the theorem into a much weaker proposition.

We will show that Melvin's claim does not hold in general; that it is true if, and only if, both countries have identical relative factor endowments—a definitely uninteresting case. The example from which Melvin generalizes is often a *possibility* when endowment ratios differ, and this possibility alone is sufficiently damaging to the Heckscher-Ohlin theorem to merit comment. But, as we shall see, the damage is much less than Melvin would have it.

I. Notation and Assumptions

We assume a world in which each country has the (same) technology to produce three goods, and each has a fixed endowment of two factors, labor and capital. For the production functions we write:

$$X_1 = F_1(K_1, L_1) \tag{1}$$

$$X_2 = F_2(K_2, L_2) \tag{2}$$

$$X_3 = F_3(K_3, L_3) \tag{3}$$

These are assumed to be homogeneous of degree one, and to exhibit positive but decreasing marginal productivity in each factor. Both factors are assumed to be fully employed. Thus,

$$L = L_1 + L_2 + L_3 \tag{4}$$

$$K = K_1 + K_2 + K_3 \tag{5}$$

for each country. By k we denote a country's endowment capital-labor ratio, K/L.

II. The Trade Pattern With Two Countries

In the three-good, two-factor, two-country case, the indeterminacy in production and trade appears when the terms of trade and each country's endowment ratio are such that it is possible for each country to produce all three goods. Taking the isoquant approach, the situation can be depicted as in Figure 1. Here, the isoquant of unit value output (in terms of, say, good 1) for each good is tangent to the unit isocost line ab, and the endowment ratio rays, k^a and k^b, lie in the cone coe. Except for the endowment rays, the same diagram serves for both countries since the production functions are assumed to be identical.

Suppose in equilibrium country A produces good 2 on a scale indicated by point f, and country B produces good 2 on a scale indicated by point h. Then we can determine by parallelogram construction that country A will necessarily produce goods 1 and 3 on a scale indicated by points g and j, respectively, and country B will necessarily produce goods 1 and 3 on a scale indicated by points m and n, respectively.[1] Assuming both countries have similar tastes, it is clear country B will export goods 2 and 3 and import good 1.[2] Thus, the depicted situation is

* Assistant professor of economics, University of Dayton. This paper was written while the author was a National Science Foundation Graduate Fellow at the University of Oregon. He thanks Chulsoon Khang for helpful suggestions.

[1] The parallelogram is the geometrical representation of vector addition. To satisfy the factor employment assumption in Figure 1 for, say, country A we must have $\overline{of}+\overline{og}+\overline{oj}=\overline{op}$, where the bar denotes a vector. Given $\overline{op}$ and $\overline{of}$ we find $\overline{oq}$ by construction such that $\overline{of}+\overline{oq}=\overline{op}$. We then find $\overline{og}$ and $\overline{oj}$ such that $\overline{og}+\overline{oj}=\overline{oq}$.

[2] By "similar tastes" we mean demand conditions do not differ enough at any relevant terms of trade to cause the larger producer of any good to import it. Melvin

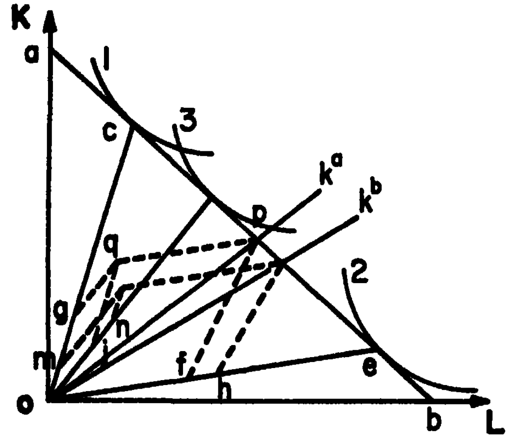


FIGURE 1

one in which country B is exporting the labor intensive good and importing the capital intensive good. We have, then, constructed a counterexample to Melvin's claim that the country which exports the labor intensive good will also export the capital intensive good. By manipulating the diagram one can convince oneself that the counterexample *can always* be constructed as long as $k^a \neq k^b$ and both are in the cone *coe*. Thus, $k^a = k^b$ is necessary for Melvin's proposition.

Suppose both countries have the same endowment ratio. Figure 2 illustrates this situation. Obviously, whichever country produces at point *a* and, hence, exports good 2, will also produce at point *c* and export good 1. Clearly, then, $k^a = k^b$ is a sufficient condition for Melvin's proposition. Therefore, we have shown that identical relative factor endowments is a necessary and sufficient condition for Melvin's proposition. A corrected version of his claim is: if both countries have the same endowment ratio, then country X exports the labor (capital) intensive good if and only if country X exports the capital (labor) intensive good.

We find no error in Melvin's construction of the production possibility surface. It appears he simply overlooked those possible trade patterns contrary to his claim but consistent with the standard Heckscher-Ohlin theorem, for our counterexample is easily illustrated with his own diagram (p. 1262) which we reproduce in Figure 3.

Here the goods are numbered such that good 2 is labor intensive and good 1 is capital intensive. Keeping Melvin's consumption point T, we have selected different production points S and R. We see the country producing at R produces goods 1 and 2 in quantities represented by points *b* and *d*, respectively, whereas the similar points for the country producing at S are *a* and *c*. Therefore, the country producing at S exports the labor intensive good and imports the capital intensive good, contrary to Melvin's claim.

If both countries have the same endowment ratio, JH and $J'H'$ will coincide and contain the consumption point T, in which case both countries could produce at T and trade would be unnecessary. For this reason we feel this special case is of little importance.

III. The Heckscher-Ohlin Theorem

On the basis of his claim Melvin turns to a weak form of the Heckscher-Ohlin theorem. According to this version a country's export bundle will be intensive, relative to its import bundle, in that factor which the country holds in relative abundance. This statement is true whether endowment ratios are

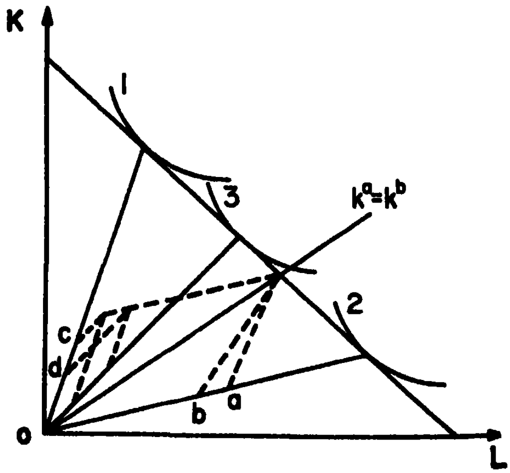


FIGURE 2

avoids explicit demand assumptions, but assumes (p. 1254) countries are identical except for their capital stocks.

equal or not. If they are equal ($k^a = k^b$), Melvin's claim holds and his Heckscher-Ohlin theorem tells us the capital-labor ratio of all traded bundles is the same. If endowment ratios are unequal the theorem is still true, but because Melvin's claim is untrue the theorem is unnecessarily weak.

We will now develop a stronger version of the Heckscher-Ohlin theorem to apply to the indeterminancy case when endowment ratios are different. In this situation there are two possibilities: either 1) each country must produce either the capital intensive or labor intensive good, or alternatively 2) one country's endowment ratio is equal to the capital-labor ratio of the good of intermediate intensity and it produces only that good. This second possibility would be rather unusual and the pattern of trade is easily resolved, for obviously the other country will export both the capital intensive and labor intensive goods.

Figure 4 gives an example of the first possibility. Here both countries must produce good 2. If country *A* were to produce only goods 2 and 3, they would be produced on a scale represented by points *a* and *n*, respectively. Similarly, points *b* and *j* are the appropriate points if only goods 2 and 1 are produced. Thus the interval *ab* represents country *A*'s range of possible outputs of

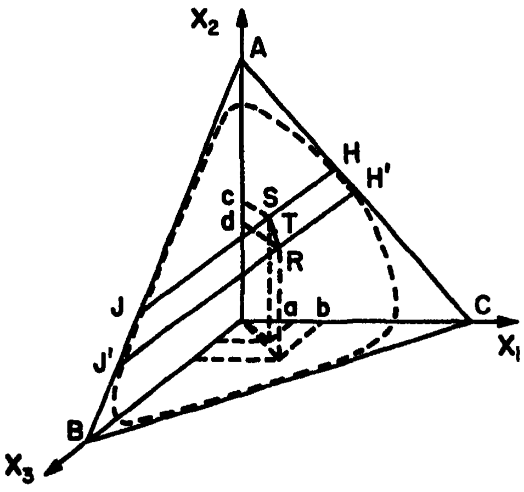


FIGURE 3

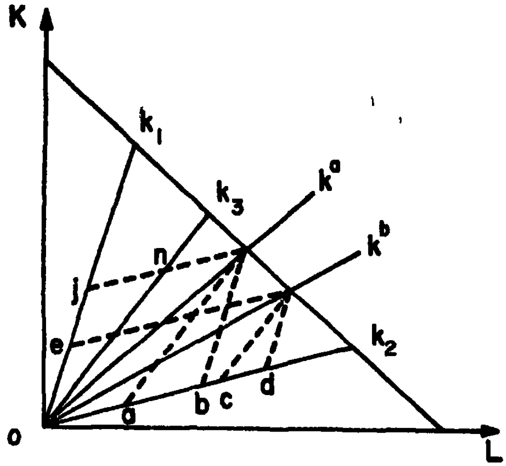


FIGURE 4

good 2. For country *B* this interval is *cd*. Since the intervals do not overlap country *B* must export good 2. It is always possible to draw k^a and k^b between k_3 and k_2 such that *ab* and *cd* do not overlap. For good 1, *oe* and *oj* represent the production ranges of countries *B* and *A*, respectively. Since $oe < oj$ country *A* is more likely to export good 1 than is country *B*. If k^a and k^b were closer together in the same price situation *ab* and *cd* would overlap and allow the possibility of country *A* exporting good 2. Also, the ratio oe/oj would be larger, thus decreasing the likelihood of country *A* exporting good 1.

Another example of the first possibility is the situation where k^a and k^b are on opposite sides of k_3. In this case it turns out that if k^a and k^b are sufficiently different the production intervals of the two countries will not overlap for either good 1 or good 2, and the pattern of trade is that predicted by the standard Heckscher-Ohlin theorem. We leave the construction of this example to the reader.

Generalizing from examples of this type, we state our theorem for the indeterminancy case with $k^a \neq k^b$ and excluding possibility 2) above. As $|k^a - k^b|$ increases, the probability of each country exporting the good using intensively that country's abundant factor increases; if both countries must produce *different* goods, there is a value of $|k^a - k^b|$

above which each country must export the good using intensively its abundant factor; if both countries must produce the *same* good, there is a value of $|k^a - k^b|$ above which that good must be exported by the country abundant in its intensively used factor.

Our theorem contains significantly more information than Melvin's, yet allows for the contradiction of the standard theorem which he points out. For sufficiently different endowment ratios (remaining within the indeterminancy case) it gives the standard result that each country will export the good using intensively that country's abundant factor.

REFERENCE

J. R. Melvin, "Production and Trade with Two Factors and Three Goods," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1249–68.

Production Indeterminacy with Three Goods and Two Factors: Reply

By JAMES R. MELVIN

Let me begin by agreeing in part with Douglas Stewart's argument. My claim that whenever all goods are traded, the country exporting the labor intensive good will also export the capital intensive good, does not hold in general. While the case I refer to is possible, I am, as Stewart has shown in his Section II, certainly guilty of a careless and incorrect generalization.

At the same time, Professor Stewart is guilty of much the same kind of error, for his statement that my claim "... is true if, and only if, both countries have identical relative factor endowments ..." (p. 241) is incorrect. Indeed one would be tempted to interpret this remark as a slip of the pen except for the fact that two similar statements are made subsequently. After correctly presenting a counterexample to my claim he says "Thus, $k^a=k^b$ is necessary for Melvin's proposition," (p. 242) and later "Therefore, we have shown that identical relative factor endowments is a necessary and sufficient condition for Melvin's proposition" (p. 242). Stewart himself presents a counterexample to all these statements when he says "The example from which Melvin generalizes is often a *possibility* when endowment ratios differ, ... " (his italics). Obviously if my claim is a possibility when endowment ratios differ, equal endowment ratios cannot be a necessary and sufficient condition for my claim. That the situation I have described is possible can be seen from Stewart's Figure 4. If endowments are such that *ab* and *cd* overlap, then it is possible for either country to export both the capital intensive and the labor intensive good while importing the intermediate one.

From my analysis I drew the conclusion that for the three-good, two-factor case, the traditional statement of the Heckscher-Ohlin Theorem must take the somewhat weaker form that a country will export a bundle of goods which is intensive in the factor with which that country is relatively well endowed. In light of the fact that my example is still a possibility, I see no reason to change my conclusion. Stewart's argument has shown that the more traditional trade pattern is possible, but that is quite consistent with my reformulation of the theorem.

In his Section III, Stewart attempts to present a stronger version of the theorem than the one I presented. Unfortunately his analysis suffers from his failure to appreciate the fact that a possibility other than the two he mentions can exist; namely that a country may export both the capital intensive and the labor intensive goods while importing the intermediate one. Furthermore, while I must confess that I have had some difficulty understanding his theorem,[1] it does not seem to add much to our knowledge. If my interpretation of the spirit of his theorem is correct it states that the more different are the endowments of the two countries, the more likely is a country to export the commodity which is intensive in the factor with which that country is relatively well endowed. While this is certainly correct it is not particularly surprising, and unless precise conditions can be derived which will tell us when it is *not* possible for one country to export both the capital intensive and the labor intensive goods, we

[1] I do not understand what it means to say that "... both countries must produce *different* goods ..." or that "... both countries must produce the *same* good, ..." (p. 244). Why *must* they do either of these? He further says that "... if both countries must produce the *same* good, there is a value of $|k^a-k^b|$ above which that good must be exported by the country abundant in its intensively used factor." But what happens, as seems likely, if the "same good" turns out to be the intermediate one?

do not seem to have advanced much beyond my statement of the theorem.

In conclusion, it should be pointed out that the discussion surrounding Stewart's Figure 3 is incorrect, for since the origins for the two countries are not the same, the quantities produced for the country producing at *R* cannot be shown on this diagram, and the quantity comparisons he draws are therefore inappropriate.

IN MEMORIAM

JACOB VINER

1892–1970

Jacob Viner, one of the great teachers and scholars of economics of our time and all times and former president (1939) of the American Economic Association, died in Princeton, New Jersey on September 12, 1970 at the age of 78.

Born in Montreal, Canada on May 3, 1892, Professor Viner attended McGill University (B. A., 1914) and then did his graduate work at Harvard University (M.A., 1915; Ph.D., 1922). His doctoral dissertation, written under Professor Frank Taussig, received the David Ames Wells Prize, and was published in 1924 as *Canada's Balance of International Indebtedness*. This book, which quickly established his reputation as an economic theorist of the first rank, not only extended the pure theory of international adjustment but was also a pioneer venture in combining rigorous theory with the thorough empirical testing of theoretical propositions.

In 1916, Viner joined the faculty of economics of the University of Chicago, where he remained until 1946, when he moved to Princeton University. He retired from teaching in 1960, but continued active intellectual work as an Emeritus Professor at Princeton and as a permanent member of the Institute for Advanced Study.

The American Economic Association awarded him its highest honor, the Francis A. Walker Medal, in 1962. Thirteen universities in the United States and abroad recognized his professional achievements with honorary degrees.

Viner was active in public affairs, particularly from 1934 to 1939, when he was special assistant to the Secretary of the Treasury, and when he played an important role in developing a research staff at the Treasury and in shaping national economic policy. In addition, he was special adviser to the U.S. Tariff Commission and U.S. Shipping Board during World War I; consultant to the Department of State 1943 to 1952; consultant to the Board of Governors of the Federal Reserve System; member of the Board of Directors of the National Bureau of Economic Research.

Yet these public activities were very much a side line. Viner's primary interests and activities were academic. Like his own mentor, F. W. Taussig, Viner was a great teacher. His basic course in Price Theory at the University of Chicago, which many students took in their first quarter of graduate study, was a deep intellectual experience. Some found it forbidding and developed a fear of the man and the subject. But to the abler students, it opened a new world. It gave them a feeling for the subtlety, power, and appeal of pure economic theory. In Viner's hands, economic theory was not a set of formal abstract propositions; it was a set of tools, to be constructed with care and the utmost attention to logical rigor, but to be judged primarily by its usefulness in understanding and interpreting important economic phenomena. He presented economic theory as, in Marshall's words, "an engine of analysis." And he presented it with verve and color, making it an exciting and controversial subject. He had few peers for quickness of mind and tongue or ability to grasp new ideas or to spot and expose fallacies.

Like Taussig also, he was a great editor. He edited or coedited the *Journal of Political Economy* for 18 years, from 1928 to 1946, keeping it in the forefront of the professional economic journals of the world. He set and maintained the very highest standards of integrity in editing, and displayed a catholicity and tolerance that made the *JPE* a widely ranging journal containing contributions embodying different approaches and reflecting different schools of thought.

Viner's scholarly influence was exerted in many areas: In price theory, where his article

on cost curves has been justly famous as a major improvement in Marshall's analysis of the relation between costs and supply; in the history of economic thought, where his incisive review of the enormous mercantilist pamphlet literature led to the rehabilitation of Adam Smith's judgment on mercantilism as a theory; in studies of the balance of power and the economic aspect of imperialism, which stimulated extensive further work by his students as well as other scholars.

His major scholarly impact, however, was unquestionably in international trade. In addition to his doctoral thesis, he published a number of seminal works, including *Dumping* (1922), *Studies in the Theory of International Trade* (1937), and *Customs Unions* (1950). His *Studies* combined his special interest in trade with his deep and abiding interest in the history of economic thought. It remains the definitive work on many aspects of the development of both monetary theory and international trade theory. His small book on *Customs Unions* introduced a fresh analysis of an ancient subject that has dominated further discussion of the problem.

Viner was a man of wide culture and interests, a fascinating conversationalist, warm friend, devoted father and husband, an ever-ready source of advice, and help to a legion of present and former students, colleagues and friends.

ERRATA

The Optimum Lifetime Distribution of Consumption Expenditures: A Correction

By Lester C. Thurow

In an iteration of comments and responses, the September 1970 *Review* inadvertently published my original response to the comments of Brian Motley and Samuel A. Morley on "The Optimum Lifetime Distribution of Consumption Expenditures" rather than my final response (pp. 744–45). The last three paragraphs of the published reply are incorrect. They should be replaced by the following three paragraphs.

If the quantitative magnitudes of the estimates are to be explained by this factor, however, there must be a very peculiar distribution of lifetime incomes over different zero-saver age groups. The lifetime income of a zero-saver man 60 years old would have to be just half of the lifetime income of a zero-saver man 40 years old if this factor were to explain savings behavior by itself. (Assuming that economic growth leads to a real rate of growth of 2 percent per year in family income, the average lifetime income of a 40 year old is 44 percent higher than that of a 60 year old.) At the other end of the scale, the lifetime income of a zero-saver man 20 years old would have to be less than the lifetime income of a zero-saver man 40 years old to explain the estimated results. I can swallow the former but not the latter.

Given identical and homothetic preferences, I agree with Motley and Morley that equation (1) is the correct estimating equation. *It is exactly the equation which I used.* The problem is to measure the general average lifetime income ($\overline{W}$) and the average lifetime income of the zero saver (W^*). In

my paper the definitions were as follows:

$$(1) \qquad \overline{W} = \sum_{t=1}^{T} \overline{Y}_t$$

$$(2) \qquad W^* = \sum_{t=1}^{T} Y_t^* = \sum_{t=1}^{T} C_t^*$$

where:

$Y_t^* = C_t^*$ for the zero-saver
$\overline{Y}_t$ = income of average t-year old household.

I also agree that both are at best only approximations to actual lifetime incomes. Fortunately some of the errors in such an approximation may disappear since the same types of biases are apt to appear in both $\overline{W}$ and W^*, and only the ratio of $\overline{W}$ to W^* is used in the analysis.

I find the Motley-Morley method of calculating optimum life-time incomes interesting, but I am puzzled as to why their model becomes the Thurow technique when it yields incorrect results. They show that equation (7) yields a convex path when the normal income path is concave and the true optimal path is postulated as constant. I agree with the conclusions, but equation (7) is not the estimating equation which I used. It is even labelled an alternative model. Thus, I can hardly consider the incorrect results a critique of my estimating equation.

A Growth Model of International Direct Investment

By Hans Brems

In the June, 1970 issue of this *Review* on p. 328, I assumed entrepreneurs to ignore their own influence upon national money income Y_i and hence upon the volume of savings $(1-c_i)Y_i$ they were allocating between parent firm and foreign subsidiary. This assumption was unnecessary. Use (92) through (95) and (100) through (103) to see that $g_{P11}+g_{X11}=G+g_{P21}+g_{X21}$, hence

$$Z_1(\tau) = [r_1 - (g_{P11} + g_{X11})]\zeta_1(\tau)$$

Insert this into (60) written for time τ; then take the partial derivative of (60) with respect to $S_{11}(\tau)$

$$\frac{\partial Y_1(\tau)}{\partial S_{11}(\tau)} = [r_1 - (g_{P11} + g_{X11})] \frac{\partial \zeta_1(\tau)}{\partial S_{11}(\tau)}$$

Consequently, maximizing $\zeta_1(\tau)$ by taking the partial derivative $\partial\zeta_1(\tau)/\partial S_{11}(\tau)$ and setting it equal to zero also makes the partial derivative $\partial Y_1(\tau)/\partial S_{11}(\tau)$ zero.

Thus all solutions of my model hold whether or not entrepreneurs ignore their own influence upon national money income and savings.

NOTES

An amendment to the bylaws adopted by mail ballot last year provides for nomination by petition to the offices of President-elect, Vice President, and member of the Executive Committee. To be valid, petitions must reach the Secretary by August 1. Petitions for President-elect must be signed by 6 percent of the members; for other offices by 4 percent of the members. The exact text of the new bylaw will be published in the May issue of this *Review*.

The names of nominees selected by the Nominating Committee will be published in the June issue of this *Review*. Members wishing to have this information earlier may contact the Secretary by mail or telephone at any time after April 1. The Secretary will make every effort to facilitate efforts to obtain the requisite number of signatures on petitions. In particular, address labels for all members will be furnished at cost.

The Economics Institute is sponsored by the American Economic Association for foreign students of economics and agricultural economics beginning graduate work in the United States.

The 14th session is to be held from June 16 to August 18, 1971 at the University of Colorado. Information and application forms may be obtained from the Director of the Economics Institute, University of Colorado, Boulder, Colorado 80302.

The American Economic Association Nominating Committee for 1971 was listed on page 1021 of the December 1970 *Review*. Karl Gregory, Oakland University should be added to this list. The omission of his name was an error.

The American-Scandinavian Foundation, the Council for European Studies, and the Society for the Advancement of Scandinavian Study will cosponsor a one-day symposium/convention devoted to "Social Science Research in Scandinavia." The symposium, to be held in conjunction with the annual meeting of S.A.S.S. in Lexington, Kentucky next May, will focus on the problems and methodology of scholarly research in the Nordic countries in the fields of history, political science, economics, sociology, and other closely related fields.

The academic director of the symposium will be Stephen Bland, associate professor of political science, University of Pittsburgh and executive director, Council for European Studies. The administrative director is Gene G. Gage, executive secretary of The American-Scandinavian Foundation and secretary-treasurer of The Society for the Advancement of Scandinavian Study. For further information, contact: Mr. Gage, *ASF*, 127 East 73rd Street, New York, N.Y. 10021, Telephone 212+TR9-9779.

The 1971 annual meeting of the New York State Economics Association (NYSEA) will be held April 16–17 at Rensselaer Polytechnic Institute, Troy, New York. Officers of NYSEA for 1970–71 are: President: Edwin J. Holstein, Rensselaer Polytechnic Institute; Vice President: David A. Martin, State University College at Geneseo; Secretary-Treasurer: Nancy R. Auster, State University of New York Agricultural and Technical College at Canton; Editor: Daniel Feinberg, New York City Community College.

New data for researchers on labor force behavior and work attitudes: First results of the early surveys made for a five-year study of the labor market experience, characteristics, and work attitudes of four groups of 5,000 people each are being made available on data tapes as well as in print. Data for the four groups: men 45–59 years old; women 30–44 years old; and young men and women 14–24 years old; come from annual surveys funded by the Labor Department's Manpower Administration. The final surveys will be conducted in 1973. The surveys, beginning in 1966 for the two groups of men, 1967 for the older women, and 1968 for the young women, cover: Current employment, work history, education and training, migration, earnings, income, family status, health, work attitudes; plus information of special significance for each group. For example, retirement plans of older men, and the costs of child care and transportation for employed older women.

Data tapes may be obtained at cost through the Chief of the Demographic Surveys Division, Bureau of the Census, U.S. Department of Commerce, Washington, D.C. 20233. Tape specifications were published in late 1970. The tapes will not identify individuals, in line with the Census Bureau's policy of confidentiality.

Printed reports may be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Reports now available cover the first two studies of older men, the first study of young men, and the first study of older women. Order as Manpower Research Monographs Nos. 15-I, 15-II, 16-I, and 22-I, respectively.

A visiting lecturer program in statistics has been organized for the eighth successive year. The program is sponsored jointly by the principal statistical organizations in the United States, the American Statistical Association, the Biometric Society, and the Institute of Mathematical Statistics. The National Science Foundation provides partial financial support. Leading teachers and research workers in statistics from universities, industry, and government, have agreed to participate as lecturers. Lecture topics include subjects in experimen-

tal and theoretical statistics, as well as in such related areas as probability theory, information theory, and stochastic models in the physical, biological, and social sciences.

The purpose of the program is to provide information to students and college faculty about the nature and scope of modern statistics, and to provide advice about careers, graduate study, and college curricula in statistics. Inquiries should be addressed to: Visiting Lecturer Program in Statistics, Department of Statistics, Southern Methodist University, Dallas, Texas 75222.

The National Science Foundation will sponsor the following College Teacher Summer Programs for 1971:

For teachers of economics: June 7–July 30. Mathematics for economics, economic theory (price and national income), managerial economics, and stabilization and growth policies. Dr. William H. Wesson, Jr., department of economics, University of South Carolina, Columbia, South Carolina 29208.

June 27–July 24: Urban economic growth and structure, human resources in urban economies, and the public sector of urban economies. Dr. Henry Levin, department of economics, Stanford University, Stanford, California 94305.

June 21–August 13: Micro- and macroeconomic analysis and its relationship to the principles courses; workshop on problems of junior college teachers of introductory economics. Dr. Jay W. Wiley, department of economics, Purdue University, Lafayette, Indiana 47907.

For teachers of introductory economics: June 7–July 16. Current issues of economic policy; and the development and teaching of a policy oriented, basic course in economics. Dr. Ewing P. Shahan, department of economics and business administration, Vanderbilt University, Nashville, Tennessee 37203.

The 1971 annual meeting of the Association of Social and Behavioral Scientists (formerly the Association of Social Science Teachers) will be held April 22 and 23 at the Holiday Inn Downtown, Montgomery, Alabama. The theme is "Environment: The Ghetto and Beyond." Program chairman is Russell Stockard, and inquiries may be addressed directly to him at the department of geography, Southern University at New Orleans, New Orleans, Louisiana.

The Public Policy Program at Harvard University is conducting an open competition to secure case studies of the application of economic and other analytic methods to real public policy problems. All completed works not previously published are eligible. Studies that have been employed in a policy-making context are particularly sought. There is no limit on length; extensive original data may be included.

Submissions will be reviewed promptly by a committee under the direction of Professors Raiffa, Schelling, and Zeckhauser. Selected entries will receive informal publication by the Public Policy Program. They will be used in our teaching program and will be distributed to other interested readers. The authors of selected entries will maintain all rights and copyright privileges, and will be awarded a $100 honorarium.

Manuscripts will be returned to authors on request. For curriculum planning purposes, it is important that entries be received not later than June 30, 1971. They should be sent to R. Zeckhauser, Littauer Center M-41, Harvard University, Cambridge, Massachusetts 02138.

Summer *NSF* Institute in management science and operations research for college professors in management science, operations research, applied mathematics, economics, health and hospital administration to be held June 14 through July 16, 1971 at the University of Colorado, Boulder. Participants receive stipend and travel expenses.

For further information, correspond with: Dr. Donald R. Plane, Co-Director, Summer Institute, Division of Management Science, Business Building, University of Colorado, Boulder, Colorado 80302.

Deaths

Elmer C. Bratt, professor of economics emeritus, Lehigh University, Nov. 9, 1970.

Alfred H. Conrad, professor of economics, City College of New York, Oct. 18, 1970.

Howard E. Dubner, assistant professor of economics, University of Miami, Oct. 9, 1970.

Everett D. Hawkins, professor of economics, University of Wisconsin, Aug. 31, 1970.

Milton S. Heath, professor of economics emeritus, University of North Carolina, Chapel Hill, Sept. 8, 1970.

John B. Lansing, chairman, department of economics, University of Michigan, Sept. 8, 1970.

Ben S. Seligman, professor of economics and director, Labor Relations and Research Center, University of Massachusetts, Oct. 23, 1970.

Retirements

C. Richard Creek, department of economics, Colorado State University, Sept. 1970.

John M. Frikart, professor of economics, University of Arizona, June 1971.

Kent T. Healy, department of economics, Yale University, July 1, 1970.

Ross Milner, professor of agricultural economics, Ohio State University, July 31, 1970.

Arthur H. Reede, professor of economics, Pennsylvania State University, Sept. 1970.

Frederick G. Reuss, department of economics, Goucher College.

Josef Soudek, professor of economics, City University of New York, Queens College, Sept. 1970.

Visiting Foreign Scholars

Knolly Barnes, University of West Indies: visiting scholar, department of economics, University of Nebraska.

Michael Bruno, Hebrew University: visiting professor, department of economics, Massachusetts Institute of Technology.

Frits J. de Jong, University of Groningen: National Science Foundation Senior Foreign Scientist Fellowship, Florida State University, 1970–71.

Gunnar Floystad, Norwegian School of Economics and Business Administration: visiting scholar, department of economics, University of Michigan, 1970–71.

Frank H. Hahn, London School of Economics: visiting professor, department of economics, Massachusetts Institute of Technology.

Roy Harrod, Oxford University: visiting professor of economics, University of Maryland, spring 1971.

Terence W. Hutchison, University of Birmingham: visiting professor of economics, Dalhousie University, fall 1970.

Sven-Ake Johansson, Royal Institute of Technology, Stockholm, Sweden: visiting scholar, Graduate School of Business Administration, University of California, Los Angeles, Oct. 1970.

S. S. Johl, Punjab Agricultural University, India: visiting professor of economics, Ohio State University, Aug. 1970–Aug. 71.

Yoav Kislev, Hebrew University of Jerusalem: visiting research associate, department of economics, Yale University, 1970–71.

James Mirrlees, Nuffield College, England: visiting professor, department of economics, Massachusetts Institute of Technology.

Carl J. Norstrom, Norwegian School of Economics and Business Administration: visiting scholar, department of economics, University of Michigan, 1970–71.

Seiji Sakuira, Hokkaido University: visiting scholar, department of agricultural economics, Cornell University.

Hirofuma Shibata, York University, England: visiting associate professor of economics, University of Maryland, 1970–71.

Yoshio Shimizu, Konan University, Kobe, Japan: visiting scholar, department of economics, University of Colorado, 1970–71.

Baran M. Tuncer, University of Ankara: visiting research associate, department of economics, Yale University, 1970–71.

A. J. W. van de Gevel, Tilburg University: visiting assistant professor of economics, Northern Illinois University, 1970–71.

Henricus P. A. van Roosmalen, Institute of Social Studies: visiting associate professor of business administration, University of New Hampshire, fall 1971.

Promotions

Richard J. Agnello: assistant professor of business and economics, University of Delaware, June 1970.

R. G. Akkihal: associate professor of economics, Marshall University.

Thaine H. Allison, Jr.: assistant professor of economics, Central Washington State College.

Nancy R. Auster: associate professor of economics, State University of New York, Canton.

Robin Barlow: professor of economics, University of Michigan.

Merrill J. Bateman: professor of economics, Brigham Young University.

Paul T. Bechtol: professor of economics, Colorado College.

Robert A. Behren: associate professor of economics, Brooklyn College.

William C. Bonifield: associate professor of economics, Wabash College.

Eleutherios N. Botsas: associate professor of economics and management, Oakland University, July 1970.

Patricia F. Bowers: associate professor of economics, Brooklyn College.

Charles W. Bullard: professor of economics, University of North Dakota.

Winston Chang: associate professor of economics, State University of New York, Buffalo, Sept. 1970.

Allan R. Cohen: associate professor of business administration, Whittemore School of Business and Economics, University of New Hampshire.

Jerry L. Dake: associate professor of corporate finance, College of Industrial Management, Georgia Institute of Technology.

Robert H. Deans: associate professor of economics, Temple University.

Lawrence P. Donnelley: assistant professor of business and economics, University of Delaware, June 1970.

Robert B. Ekelund: associate professor of economics, Texas A&M University.

Edward G. Emerling: professor of economics, St. Bonaventure University.

John T. Etheridge: associate professor of industrial management, College of Industrial Management, Georgia Institute of Technology.

Irwin Feller: associate professor of economics, Pennsylvania State University.

Jay D. Forsyth: assistant professor of business administration, Central Washington State College.

Wolfgang W. Franz: assistant professor of economics, Central Washington State College.

A. Myrick Freeman III: associate professor of economics, Bowdoin College, fall 1970.

Bruce L. Gensemer: associate professor of economics, Kenyon College.

Richard A. Goodman: associate professor, department of business administration, Graduate School of Business Administration, University of California, Los Angeles.

Richard L. Gordon: professor of mineral economics, Pennsylvania State University, July 1, 1970.

William P. Gramm: associate professor of economics, Texas A&M University.

Bennett Harrison: assistant professor of economics, University of Maryland, Jan. 1971.

Robert E. Hicks: associate professor, department of economics and finance, Florida Technological University.

C. Russell Hill: assistant professor of economics, University of Michigan.

James M. Holmes: associate professor of economics, State University of New York, Buffalo, Sept. 1970.

Teh-wei Hu: associate professor of economics, Pennsylvania State University, July 1970.

John R. Kaatz: associate professor of economics, College of Industrial Management, Georgia Institute of Technology.

Alvin K. Klevorick: associate professor of economics, Yale University, July 1970.

A. J. Kondonassis: professor of economics, University of Oklahoma, Sept. 1970.

Ronald S. Koot: associate professor of quantitative business analysis, Pennsylvania State University, July 1970.

Iwan S. Koropeckyj: professor of economics, Temple University.

Chung H. Lee: associate professor of economics, Miami University.

Albert M. Levenson: professor of economics, Queens College, City University of New York.

Daniel Lipsky: professor of economics, Brooklyn College.

Wesley Long: associate professor of economics, Pennsylvania State University.

Robert M. Lovejoy: associate professor of economics, State University of New York, Binghamton.

John J. McGowan: associate professor of economics, Yale University, July 1, 1970.

Graeme H. McKechnie: associate professor of economics, York University.

William J. McKenna: professor of economics, Temple University.

Donald McLeod: associate professor of economics, Temple University.

James B. MacQueen: professor, Graduate School of Business Administration, University of California, Los Angeles, July 1, 1970.

Fredric C. Menz: assistant professor of economics, Temple University.

Roger N. Millen: assistant professor, Graduate School of Business, Columbia University, July 1, 1970.

Charles G. Moore: assistant professor of economics, Brooklyn College.

Dennis C. Mueller: associate professor, department of economics, Cornell University, July 1, 1970.

Eugene A. Myers: professor of economics, Pennsylvania State University, July 1970.

William D. Nordhaus: associate professor of economics, Yale University, July 1, 1970.

Carl Nordstrom: professor of economics, Brooklyn College.

J. Randolph Norsworthy: associate professor of economics, Temple University.

Patrick R. O'Shaughnessy: associate professor of business administration, Central Washington State College.

Lewis J. Perl: assistant professor, department of labor economics and income security, New York State School of Industrial and Labor Relations, Cornell University.

Barry W. Poulson: associate professor of economics, University of Colorado.

Arnold H. Raphaelson: professor of economics, Temple University.

T. Ross Reeve: assistant professor of economics, Central Washington State College.

Richard D. Reimer: professor of economics, College of Wooster, Sept. 1970.

Stephen A. Resnick: associate professor of economics, Yale University, July 1, 1970.

Richard Rosenberg: assistant professor of economics, Pennsylvania State University, Jan. 1970.

George H. K. Schenck: associate professor of mineral economics, Pennsylvania State University, July 1, 1970.

Harvey Schwartz: associate professor of economics, York University.

Richard A. Seese: assistant professor of economics, John Carroll University.

Paul Seidenstat: associate professor of economics, Temple University.

Merrill K. Sharp: assistant professor, department of economics, Iowa State University.

Philip Sheinwold: professor of economics, Brooklyn College.

Mitchell Stengel: assistant professor, department of economics and Center for Urban Affairs, Michigan State University, Sept. 1, 1970.

Samuel S. Stewart: assistant professor, Graduate School of Business, Columbia University, July 1, 1970.

Joseph E. Stiglitz: professor of economics, Yale University, July 1, 1970.

Venkataraman Sundararajan: assistant professor of economics, University College, New York University.

Richard J. Trethewey: assistant professor of economics, Kenyon College.

Joan G. Walters: associate professor of economics, Fairfield University.

Martin L. Weitzman: associate professor of economics, Yale University, July 1, 1970.

Katherine M. West: associate professor of economics, Brooklyn College.

Administrative Appointments

John W. Allen. chairman, department of economics, Texas A&M University.

Robert Andrews: assistant dean, Graduate School of Business Administration, University of California, Los Angeles, Oct. 1970.

Robert L. Aronson: chairman, department of labor economics and income security, New York State School of Industrial and Labor Relations, Cornell University, July 1970.

D. A. L. Auld: acting chairman, department of economics, College of Social Science, University of Guelph, July 1, 1970.

Morton S. Baratz: director, Boston University Urban Institute.

Thomas A. Bausch: assistant dean, School of Business, John Carroll University.

Charles E. Bishop, North Carolina University: chancellor, University of Maryland, fall 1970.

David B. Brooks, U.S. Bureau of Mines: head, economic research section, Mineral Resource Branch, Department of Energy, Mines and Resources, Ottawa.

Elwood S. Buffa: associate dean, Graduate School of Business Administration, University of California, Los Angeles, July 1970.

Herbert A. Chesler: administrative officer, department of economics, University of Pittsburgh.

Young-iob Chung: chairman, department of economics, Eastern Michigan University.

Lawrence P. Cole: assistant dean, Whittemore School of Business and Economics, University of New Hampshire.

George G. Dawson: director of research and publications, Joint Council on Economic Education; managing editor, *Journal of Economic Education*, Sept. 1970.

Lynn E. Dellenbarger, Jr.: director of graduate programs in business administration and industrial relations, College of Business and Economics, West Virginia University.

John Dorsey: vice-chancellor for Business Affairs, University of Maryland, fall 1970.

James S. Earley: dean, College of Social and Behavioral Sciences, University of California, Riverside, July 1970.

Leo Fishman: chairman, department of economics, College of Business and Economics, West Virginia University.

Michael Gort: acting chairman, department of economics, State University of New York, Buffalo, 1970–71.

R. Earl Green: associate dean, College of Industrial Management, Georgia Institute of Technology.

John D. Guilfoil: director of undergraduate studies, department of economics, New York University, Washington Square.

Bernard Hall: vice-president and provost, Kent State University.

John R. Haskell: assistant dean, Whittemore School of Business and Economics, University of New Hampshire.

Robert E. Hicks: acting chairman, department of economics and finance, Florida Technological University.

Alfred E. Hofflander: vice-chairman, department of business administration, Graduate School of Business Administration, University of California, Los Angeles, Oct. 1, 1970.

O. Henry Hoversten: head, department of business administration and economics, Capital University.

Thomas S. Isaack: chairman, department of business administration, College of Business and Economics, West Virginia University.

Harald R. Jensen: acting head, department of agricultural and applied economics, University of Minnesota.

George D. Johnson: chairman, accounting and quantitative methods department, Chico State College.

Mark J. Kasoff: chairman, department of economics, Antioch College.

J. David Lages: chairman, department of economics, Southwest Missouri State College.

Charles N. Lanier: acting chairman, department of economics, University of Delaware, 1970–71.

Albert M. Levenson: associate dean, Faculty for the Social Sciences, Queens College, City University of New York.

W. Clair Lillard: director, International Programs, Central Washington State College.

Kenneth McLennan: chairman, department of economics, Temple University.

Patrick Mann: director, graduate programs in economics, College of Business and Economics, West Virginia University.

Laurence J. Mauer: acting chairman, department of economics, Northern Illinois University.

Frederic Meyers: vice-chairman, department of business administration, Graduate School of Business Administration, University of California, Los Angeles, Oct. 1, 1970.

John R. Morris: associate chairman, department of economics, University of Colorado.

John H. Niedercorn: chairman, department of economics, University of Southern California.

Martin Plotnik: acting chairman, department of economics, Slippery Rock State College, Jan. 1971.

Roger L. Ranson: chairman, department of economics, University of California, Riverside, July 1970.

William R. Reilley: chairman, department of economics and business administration, Norwich University, July 1970.

Warren L. Smith: acting chairman, department of economics, University of Michigan.

Joseph M. Thorson: associate dean, School of Social and Behavioral Sciences, West Chester State College.

Joseph L. Tryon: chairman, department of economics, Georgetown University, summer 1971.

Hugh G. Wales, University of Illinois: acting head, department of management, College of Business Administration, Roosevelt University.

Laszlo Zsoldos: acting dean, College of Business and Economics, University of Delaware, 1970–71.

Appointments

Jan P. Acton, Harvard University: staff member, economics department, The RAND Corporation, Nov. 1970.

Arjun Adlakha: research demographer, Food Research Institute, Stanford University.

Hamilton Alexander, Old Dominion University: assistant professor of economics, Virginia Commonwealth University.

Robert Ante: assistant professor of geography, department of economics, Queens College, City University of New York, 1970–71.

Sven W. Arndt, University of California, Santa Cruz: visiting assistant professor, Food Research Institute, Stanford University, spring 1971.

Enrique Arzac: assistant professor, Graduate School of Business, Columbia University, Jan. 1, 1971.

E. Dean Baldwin, University of Illinois: assistant professor of economics, Miami University.

Joseph L. Balintfy: professor of general business and finance; department of industrial engineering, University of Massachusetts, Sept. 1970.

Milton J. Bass, University of Michigan: lecturer, de-

partment of economics, Queens College, City University of New York, 1970–71.

Henry B. R. Beale: instructor, department of economics, Georgetown University, 1970–71.

Larry G. Beall, Erskine College: assistant professor, department of economics, Virginia Commonwealth University.

Burley V. Bechdolt: assistant professor of economics, Northern Illinois University.

William S. Becker, Louisiana State University: assistant professor, department of economics, Colorado College.

Larry D. Bedford: instructor, department of economics, Iowa State University.

Sanford A. Belden: assistant professor, department of agricultural economics, Cornell University.

Herman A. Berliner, City University of New York: assistant professor of economics, Hofstra University, fall 1970.

George W. Betz, University of Singapore, Malaya: associate professor of economic development, Whittemore School of Business and Economics, University of New Hampshire.

Edward R. Bleau: instructor, department of economics, Marshall University, 1970–71.

Sam H. Book: assistant professor, faculty of administrative studies, York University, Sept. 1970.

J. Patrick Bovino, Pennsylvania State University: instructor of business administration, Whittemore School of Business and Economics, University of New Hampshire, fall 1971.

Leonard R. Boyer, Kent State University: instructor of economics, Old Dominion University.

Colin I. Bradford, Jr.: senior economist, CIAP, Department of Economic Affairs, Organization of American States, Washington.

Eric Brucker, Southern Illinois University: assistant professor of business and economics, University of Delaware, Sept. 1970.

John Buttrick: professor, department of economics, York University, June 1970.

William P. Butz, University of Chicago: staff member, economics department, The RAND Corporation, Sept. 1970.

Jann W. Carpenter: associate professor of business administration, Central Washington State College.

Johanna R. Cavanaugh: lecturer, accounting department, University of Nevada, 1970–71.

Fikret Ceyhun, Wayne State University: assistant professor of economics, University of North Dakota, 1970–71.

Chau-nan Chen: assistant professor of economics, Northern Illinois University.

C. Mark Choate: assistant professor of general business and finance, University of Massachusetts, Sept. 1, 1970.

Ronald Choy, University of California, Berkeley: staff member, economics department, The New York City RAND Institute, July 1970.

David S. C. Chu, Yale University: staff member, economics department, The RAND Corporation, Oct. 1970.

Donald J. Cocheba: assistant professor of business administration, Central Washington State College.

Patricia A. Coffey: instructor, department of economics, Iowa State University.

Neal P. Cohen, University of Wisconsin: assistant professor of economics, Eastern Michigan University.

Richard V. L. Cooper, University of Chicago: staff member, economics department, The RAND Corporation, Jan. 1971.

Philip G. Cotterill, University of Illinois, Chicago Circle: assistant professor of economics, Miami University.

Larry Cox, Kansas State University: instructor of economics, Southwest Missouri State College.

James F. Crook: assistant professor, School of Business Administration, Winthrop College, fall 1970.

Alvin M. Cruze: visiting lecturer in economics, University of North Carolina, Chapel Hill.

Robert Daniels, University of Lancaster: assistant professor, department of economics, Case Western Reserve University.

Rachel Dardis, Cornell University: lecturer in economics, associate professor of home economics, University of Maryland, fall 1970.

Paul G. Darling, Bowdoin College: visiting professor of economics, McGill University, spring 1971.

William W. Davis, University of Kentucky: instructor, department of economics, Western Kentucky University, 1970–71.

Michael E. dePrano, University of Southern California: visiting professor of economics, Texas A&M University.

John Despres, University of California, Berkeley: staff member, economics department, The RAND Corporation, Sept. 1970.

James Doane, Northeastern University: assistant professor, department of economics, University of Maine.

Michael R. Dohan, California Institute of Technology: assistant professor of economics, Queens College, City University of New York, Feb. 1971.

Peter F. Drucker, New York University: professor of business economics, Claremont Graduate School, spring 1971.

Donald H. Ebbeler: assistant professor of economics, College of Industrial Management, Georgia Institute of Technology.

Linda N. Edwards, Columbia University: lecturer, department of economics, Queens College, City University of New York, 1970–71.

Donald Eilenstine, Baker University: associate professor of economics, Eisenhower College.

Doyle A. Eiler: assistant professor, department of agricultural economics, Cornell University.

Ghazi T. Farah, Wisconsin State University: assistant professor of economics, Florida Technological University.

John D. Ferguson, Brown University: assistant professor of economics, Miami University.

Sherwood M. Fine: visiting professor of economics, Washington and Lee University.

James J. Fralick: assistant professor, department of economics, Fordham University, 1970–71.

James H. Gapinski, University of New York, Buffalo: assistant professor, department of economics, Florida State University.

Michael H. Gieske: research associate, department of economics, Iowa State University.

Roy F. Gilbert, Michigan State University: visiting assistant professor of economics, Texas A&M University.

Frederic Glantz, Syracuse University: instructor of economics, Temple University.

Martin L. Gosman: assistant professor of accounting, University of Massachusetts, Sept. 1, 1970.

Harold G. Halcrow, University of Illinois: visiting professor, Food Research Institute, Stanford University, winter quarter 1970–71.

Robert E. Hall, University of California, Berkeley: associate professor, department of economics, Massachusetts Institute of Technology.

Stanley H. Hargrove: research associate, department of economics, Iowa State University.

Michael J. Hartley, Duke University: assistant professor of economics, State University of New York, Buffalo, Sept. 1970.

Joseph A. Hasson: visiting professor of economics, Acadia University, 1970–71.

Walter E. Hecox, U.S. Military Academy: assistant professor, department of economics, Colorado College.

James J. Heilbrun: associate professor, department of economics, Fordham University, 1970–71.

Julius Held: assistant professor of economics, Brooklyn College.

Richard Hellman, Small Business Administration: visiting professor of economics, University of Rhode Island, fall 1970.

George R. Henderson: instructor of economics, Miami University.

Louis Henry, Notre Dame University: assistant professor of economics, Old Dominion University.

Larry M. Hersh, Harvard University: lecturer, department of economics, Queens College, City University of New York, 1970–71.

S. Hugh High, Mississippi State College: instructor, department of economics, North Texas State University.

Roger B. Hill, University of Georgia: professor of economics, University of North Carolina, Wilmington.

George E. Hoffer, University of Virginia: assistant professor, department of economics, Virginia Commonwealth University.

Joe R. Hulett, Iowa State University: assistant professor of economics, Texas A&M University.

John A. Hutcheson: visiting assistant professor of economics, Queen's University.

Milton A. Iyoha, Yale University: assistant professor of economics, State University of New York, Buffalo, Sept. 1970.

William Jaffe: professor, department of economics, York University, June 1970.

John B. Kaye, George Washington University: lecturer, management department, University of Nevada, 1970–71.

William M. Kempey, Hofstra University: teaching associate, department of economics, Queens College, City University of New York, 1970–71.

A. Thomas King, Yale University: research assistant, Bureau of Business and Economic Research; lecturer in economics, University of Maryland, fall 1970.

Walter J. Klages, York University: associate professor of economics, Florida Technological University.

Tetteh A. (Ben) Kofi: acting assistant professor, Food Research Institute, Stanford University.

Joseph Kowalski, Wayne State University: instructor of economics, Temple University.

Theodore J. Kreps: visiting professor of economics, Central Washington State College.

Prem Kumar: assistant professor of general business and finance, University of Massachusetts, Sept. 1, 1970.

Eddy L. LaDue: assistant professor, department of agricultural economics, Cornell University.

Donald Larson, Michigan State University: assistant professor of economics, Ohio State University, Oct. 1, 1970.

Robert J. Latham: assistant professor of economics, Pennsylvania State University, Sept. 1970.

Wendy Lee, Northwestern University: assistant professor of economics, Texas A&M University.

Wayne E. Leininger: instructor in accounting, University of Massachusetts, Sept. 1, 1970.

Charles R. Link: instructor of business and economics, University of Delaware, Sept. 1970.

Jay Lowden, Jr.: instructor, division of business and economics, Nasson College, Sept. 1970.

Mark A. Lutz: assistant professor, department of economics, University of Maine.

Charles E. McConnel, Occidental College: assistant professor of economics, University of Texas, El Paso, 1970–71.

Majorie McElroy: assistant professor, department of economics, Duke University, 1970–71.

W. Terrence McGrath, University of Southern California: visiting assistant professor of economics, University of Maryland, 1970–71.

Jacob Merrieweather: associate in business, Graduate School of Business, Columbia University, Oct. 1. 1970.

Richard G. Milk: assistant professor of economics, College of Business Administration, Northeast Louisiana University.

Ira J. Miller: assistant professor, department of economics, Southern Methodist University, fall 1970.

L. Charles Miller, Jr., Haverford College: assistant professor of economics and director of internship program in urban studies, Yale University, July 1, 1970.

Stephen M. Miller: instructor, department of economics, University of Connecticut, Sept. 16, 1970.

William J. Moore, University of Oklahoma: assistant professor, department of economics, University of Houston.

Max Moszer, Bureau of Labor Statistics: professor of economics, Virginia Commonwealth University.

Roger Murray: professor of banking and finance, Graduate School of Business, Columbia University, Jan. 1, 1971.

Selma J. Mushkin: professor of economics, Georgetown University, 1970–71.

M. Ishaq Nadiri, Columbia University: professor of economics, New York University.

Lester A. Neidell: associate professor of marketing, College of Industrial Management, Georgia Institute of Technology.

Roger A. Norem: instructor, department of economics, Iowa State University.

Guy H. Orcutt, Urban Institute: professor of economics and urban studies, Yale University, July 1, 1970.

Alex Orden: professor of industrial management, College of Industrial Management, Georgia Institute of Technology.

Harry T. Oshima, University of Hawaii: visiting professor of economics, University of Pittsburgh, summer 1971.

David B. Pariser, Southern Illinois University: assistant professor of economics, University of North Dakota, 1970–71.

Robert J. Parsons: instructor, department of economics, Brigham Young University, 1970–71.

John Pisarkiewicz, Vanderbilt University: instructor, department of economics, Western Kentucky University, Aug. 1970.

Lon Polk: assistant professor of economics, Oakland University.

Daniel Primont, University of California, Santa Barbara: assistant professor of economics, Temple University.

Ronald E. Raikes: assistant professor, department of economics, Iowa State University.

Samuel D. Ramenofsky: instructor, department of economics, University of Missouri, fall 1970.

David E. Ramsett, University of Northern Iowa: associate professor of economics; director of North Dakota Center for Economic Education, University of North Dakota.

Suzanne E. Reid: assistant professor of economics, College of Business and Economics, West Virginia University.

Rudolf Rhomberg, International Monetary Fund: visiting professor of economics, University of Pittsburgh, fall/winter 1970–71.

Samuel B. Richmond: associate dean and professor, Graduate School of Business, Columbia University, Jan. 1, 1971.

M. Richard Roseman, University of Iowa: associate professor of economics, California State College, Los Angeles, Sept. 1970.

Bernard Rostker: economist, management sciences department, The RAND Corporation, Sept. 1970.

Timothy P. Roth, Texas A&M University: assistant professor, department of economics and finance, University of Texas, El Paso, 1970–71.

Dick L. Rottman, University of Missouri, Columbia: associate professor, University of Nevada, 1970–71.

Domenick Salvatore: instructor, department of economics, Fordham University, 1970–71.

Kehar S. Snagha, University of North Carolina, Charlotte: professor of economics, Old Dominion University.

Kazuo Sato, United Nations and Massachusetts Institute of Technology: professor of economics, State University of New York, Buffalo, Sept. 1970.

Linda Schumacher: instructor, department of economics, Hofstra University, fall 1970.

Abdelalem M. Sharshar, George Washington University: assistant professor, department of economics, Virginia Commonwealth University.

George W. Sheldon: instructor in general business and finance, University of Massachusetts, Sept. 1, 1970.

Robert Shishko, Yale University: staff member, economics department, The RAND Corporation, Sept. 1970.

Uma K. Sirivastava: research associate, department of economics, Iowa State University.

David J. Smyth, State University of New York, Buffalo: professor of economics, Claremont Graduate School, spring 1971.

Hugo H. Soll, Universidad Autonoma de Gualdalajara: assistant professor of economics, University of North Dakota.

John C. Soper: assistant professor, department of economics, Central Michigan University.

Philip E. Sorensen, University of California, Santa Barbara: associate professor, department of economics, Florida State University, Jan. 1971.

Brent W. Spaulding: instructor, department of economics, Iowa State University.

Samuel R. Stivers: assistant professor of industrial management, College of Industrial Management, Georgia Institute of Technology.

Donald E. Stone: associate professor of accounting, University of Massachusetts, Sept. 1, 1970.

Raymond S. Strangways, Georgia State University: professor of economics, Old Dominion University.

Martha Strayhorn, University of Wisconsin: lecturer in economics, University of Maryland, fall 1970.

Richard D. Teach: associate professor of marketing, College of Industrial Management, Georgia Institute of Technology.

William Thomas, U.S. Department of Commerce: assistant professor, department of economics, University of Houston.

G. Richard Thompson: assistant professor of economics, Florida Technological University.

Samuel L. Thorndike, Jr., University of Wisconsin, Milwaukee: assistant professor of economics, Bowdoin College, 1970–71.

Willard L. Thorp: professor of economics, University of Florida, Jan.–Mar. 1971.

John E. Tilton, Brookings Institution: assistant professor of economics, University of Maryland, fall 1970.

Subidey Togan, Johns Hopkins University: visiting assistant professor of economics, Texas A&M University.

Burke O. Trafton, American Metal Climax, Inc.: instructor of mineral economics, Pennsylvania State University.

Richard L. Tuley: instructor of economics, College of Industrial Management, Georgia Institute of Technology.

David J. Vail, Makerere University College, Uganda, Africa: assistant professor of economics, Bowdoin College, 1970-71.

Evangelina Vives: assistant professor of economics, Brooklyn College.

Robert P. Volyn, Upsala College: assistant professor of economics and business administration, Wagner College.

Gerald von Dohlen: assistant professor, Graudate School of Business, Columbia University, July 1, 1970.

Walter J. Wadycki, University of Illinois, Chicago Circle: assistant professor of economics and business analysis, Miami University.

Forrest E. Walters: associate professor, department of economics, Colorado State University.

Charles L. Weber: visiting assitant professor of economics, Valparaiso University, 1970-71.

James M. Weglin: lecturer in business administration, Central Washington State College.

David L. Weiss: lecturer in business administration, Central Washington State College.

Frederick J. Wells: visiting lecturer in economics, University of North Carolina, Chapel Hill, 1970-71.

Tom S. Witt: assistant professor of economics, College of Business and Economics, West Virginia University.

Dick R. Wittink: lecturer in business administration, Central Washington State College.

Edwin T. Wood: lecturer of business and economics, University of Delaware, Sept. 1970.

Robert A. Young: associate professor, department of economics, Colorado State University.

Jeffrey F. Zabler, University of Pennsylvania: assistant professor of economics, Wheaton College, Sept. 1970.

Leaves for Special Appointments

Dwight S. Brothers, Harvard University: program advisor in economic development and administration for East Africa, Ford Foundation, Sept. 1970-Aug. 1972.

Martin J. Davidson, North Texas State University: visiting professor, Bar-Ilan University, Israel, Jan. 1971.

William P. Dillingham, Florida State University: visiting professor, Fulbright Fellow, University of Roasrio, Argentina.

Romesh K. Diwan, Rensselaer Polytechnic Institute: visiting professor, Graduate School of Economics, Washington University.

Thomas D. Duchesneau, University of Maine: staff economist, Bureau of Economics, Federal Trade Commission.

Isaiah Frank, Johns Hopkins University: executive director, President's Commission on International Trade and Investment Policy, 1970-71.

Walter Galenson, Cornell University: professor of American history and institutions, Cambridge University, England, 1970-71.

Jerome W. Hammond: University of Minnesota-Tunisia project under contract with the University of Minnesota and U.S. Agency for International Development.

Ernst W. Kuhn, University of Nebraska: Han schochschule, St. Gallen, Switzerland.

Donald Larson, Ohio State University: Institut Building Contract, Piracicaba, Brazil, Dec. 1970-72

David F. Lean, Franklin and Marshall Colle economist, Federal Trade Commission, 1970-71.

Scott R. Pearson, Food Research Institute, Stanf University: economist, President's Commission International Trade and Investment Policy, 1970-71

Malcolm J. Purvis, chief of University of Minnes Tunisia project under contract with the University Minnesota and U.S. Agency for International Devel ment.

Rex D. Rehnberg, Colorado State University: p fessor of agricultural economics, University Nebraska-MASUA, program in Colombia, So America, 1970-72.

Lee P. Robbins, Franklin and Marshall Colle assistant director, Urban Program of the Cen Pennsylvania Consortium of Colleges, 1970-71.

Donald D. Rohdy, Colorado State University: c sultant, dean of agriculture, Virginia Polytechnic In tute, 1970-71.

Larry E. Ruff, University of California, San Die U.S. Treasury Department, Office of Tax Analy Washington 1971-72.

Anthony E. Scaperlanda, Northern Illinois U versity: visiting professor of economics, Tilburg U versity, Tilburg, The Netherlands, 1970-71.

Wilson E. Schmidt, Virginia Polytechnic Instit and State University: deputy assistant secretary research, Office of the Assistant Secretary for Int national Affairs, Department of Treasury, Washingt

Nathaniel E. Shechter, Old Dominion Univers economic consultant to Bank of Israel, 1970-71.

Bertram Silverman, Hofstra University: resea fellow, Yale University, summer and fall 1970.

Robert P. Strauss, University of North Caroli Chapel Hill: Brookings Institution Economics Po Fellow, 1970-71.

Samuel H. Talley, University of Maine: bank markets section, Division of Research and Statist Board of Governors of the Federal Reserve System

David A. Walker, Pennsylvania State Universi associate professor of quantitative business analy F.D.I.C., Washington, 1970-71.

Douglas W. Webbink, University of North Caroli Chapel Hill: Brookings Institution Economics Pol Fellow, 1970-71.

J. Hugh Winn, Colorado State University: extens economist, Instituto Tecnologico Y De Estuc Superiores de Monterrey, Mexico, 1970-71.

Tae-Hee Yoon, Dominion Bureau of Statist Government of Canada: agricultural economist, Int national Bank for Reconstruction and Developme Washington.

Resignations

Harry Allan, University of Massachusetts: Syrac University, Aug. 31, 1970.

Miloslave Bernasek, Boston University: Macqu University, New South Wales, Australia, Sept. 1971

Irwin Bernhardt, Pennsylvania State University: University of Waterloo, Ontario, June 1970.

Philip M. deMoss, University of Missouri, Kansas City: PMC College.

Muzaffer M. ErSelcuk, Purdue University, Sept. 1970.

Charles Gallagher, University of North Dakota: University of Calgary, Alberta, Aug. 9, 1970.

Lee C. Garrison, Graduate School of Business Administration, University of California, Los Angeles, July 1, 1970.

Irving Gershenberg, University of Connecticut: Makerere, Uganda.

Peter J. Ginman, Boston University: State University of New York, Geneseo, Sept. 1971.

Eila Hanni, Colorado College.

Robinson Hollister, University of Wisconsin, Madison: Princeton University, June 1970.

David M. Johnson, University of North Dakota, June 1970.

Howard P. Kitt, Hofstra University, fall 1970.

Mark W. Leiserson, Yale University: International Labor Office, Geneva, Switzerland, Sept. 1, 1970.

Edward Lorentzen, Colorado College.

James B. McCollum, College of Industrial Management, Georgia Institute of Technology.

Harry M. Markowitz, Graduate School of Business Administration, University of California, Los Angeles, July 1, 1970.

Howard Pack, Yale University: Swarthmore College, July 1, 1970.

Rein Peterson, Graduate School of Business, Columbia University, July 1970.

Michael T. Quinn, Graduate School of Business Administration, University of California, Los Angeles, July 1, 1970.

Hans O. Schmitt, University of Wisconsin, Madison: International Bank for Reconstruction and Development, Washington, June 1970.

Philip J. Schreiner, Graduate School of Business Administration, University of California, Los Angeles, July 1, 1970.

Tibor Scitovsky, Yale University: Stanford University, July 1, 1970.

Isidore Silver, University of Massachusetts: John Jay College of Criminal Justice, New York, Aug. 31, 1970.

David Singer, Hofstra University, fall 1970.

Ernst Stromsdorfer, Pennsylvania State University: University of Indiana, June 1970.

Ward Theilman, University of Massachusetts: Texaco Company, New York, Aug. 31, 1970.

Myron Uretsky, Graduate School of Business, Columbia University, June 30, 1970.

Peter Vaill, Graduate School of Business Administration, University of California, Los Angeles: University of Connecticut, July 1, 1970.

George C. Wang, University of Southern California: University of Tennessee, Aug. 1970.

Donald H. Woods, Graduate School of Business Administration, University of California, Los Angeles, July 1, 1970.

Miscellaneous

Edwin J. Holstein: president of the New York State Economics Association.

NATIONAL REGISTRY FOR ECONOMISTS

The National Registry for Economists was established in January, 1966, to provide a centralized nationwide clearinghouse for economists on a year-round basis. It is located in the Chicago Professional Placement Office of the Illinois State Employment Service and is staffed by experienced placement personnel, operating under the guidance and direction of Regional and National Bureau of Employment Security Professional Placement officials, and in cooperation with the American Economic Association. It is a free service. There are no registration, referral, or placement fees. Application and order forms used in the Registry are available upon request from the: National Registry for Economists, Professional Placement Center, 208 South La Salle Street, Chicago, Illinois 60604.

AMERICAN ECONOMIC ASSOCIATION

VACANCIES AND APPLICATIONS

The Association is glad to render service to applicants who wish to make known their availability for positions in the field of economics and to administrative officers of colleges and universities and others who are seeking to fill vacancies in the field of economics.

The officers of the Association take no responsibility for making a selection among the applicants or following up the results. The Secretary's office will merely afford a central point for forwarding inquiries, and the *Review* will publish in this section a brief description of vacancies announced and of applications submitted (with necessary editorial changes). Since the Association has no other way of knowing whether or not this section is performing a real service, the Secretary would appreciate receiving notification of appointments made as a result of these announcements. Those submitting such announcements have the option of publishing either name and address or a key number in the listing. Inquiries about a listing with a key number should refer to it specifically and be mailed to the Secretary's office. Resumes and application blanks are not supplied by the American Economic Association. The Association will only forward inquiries and resumes to the proper party for their consideration. Deadlines for the four issues of the *Review* are January 1, April 1, July 1, and October 1.

Communications should be addressed to: The Secretary, American Economic Association, 809 Oxford House, 1313 21st Avenue South, Nashville, Tennessee 37212.

Vacancies

Fishery economists: Wide variety of economics research, ranging from international agreements, quotas, and tariffs to price analyses, business management of firms, cost-benefit analyses, and whole field of the economics of natural resources. Positions are in the federal Civil Service at Grades GS-9 ($9,320) to GS-14 ($18,531). Basic requirements are Ph.D. or master's in economics or agricultural economics; training in international, natural resource, and/or quantitative economics would be helpful. Positions are located at the University of Maryland, in Washington, D.C., and field positions. Civil Service Commission Form 171 (Application for Federal Employment) should be sent to: Personnel Office, Bureau of Commercial Fisheries, U.S. Department of the Interior, 18th and C Street, N.W., Washington, D.C., 20240.

Economist: Opening in the Department of State Planning. Master's degree in economics and six years of experience required. Additional graduate study may be substituted for two years of the required experience. Background in quantitative research methods desirable. Starting salary $13,739 effective September 1, 1970, with a maximum of $18,049 reached in six years. Write: Vladimir Wahbe, Secretary of State Planning, Department of State Planning, 301 West Preston Street, Baltimore, Maryland, 21201.

THE AMERICAN ECONOMIC REVIEW

June 1971

VOLUME LXI, NUMBER 3, PART I

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

Articles

Communications

Optimal Taxation and Public Production II: Tax Rules

By Peter A. Diamond and James A. Mirrlees*

In Part I of this paper which appeared in the March 1971 issue of this *Review*, we set out the problem of using taxation and government production to maximize a social welfare function. We derived the first-order conditions, and considered the argument for efficiency in aggregate production. Here in Part II we consider the structure of optimal taxes in more detail. Part I contained five sections, and Part II begins at Section VI. In the sixth and seventh sections we consider commodity taxation in one- and many-consumer economies. In the eighth section we consider other kinds of taxes; and in the ninth, public consumption. In the tenth section we consider a rigorous treatment of the problem, giving a sufficient condition for the validity of the first-order conditions. To begin, we shall restate the notation and basic problem.

Notation

p	producer prices
q	consumer prices
t	taxes $(t=q-p)$
$x^h(q)$	net demand by consumer h (incomes are assumed to equal zero) $h=1, 2, \ldots, H$
$u^h(x^h)$	utility function of consumer h
$v^h(q)$	indirect utility function of consumer h $v^h(q)=u^h(x^h(q))$
$X(q)$	aggregate net demand $X(q)=\sum_h x^h(q)$
$U(x^1, \ldots, x^H)$	social welfare function
$V(q)$	indirect social welfare function $V(q)=U(x^1(q), \ldots, x^H(q))$
$W(u^1, \ldots, u^H)$	special case of an individualistic social welfare function, assumed for some of the analysis below.

With this notation before us again, we can restate the welfare maximization problem as that of selecting q to

$$\text{(33)}\qquad \begin{array}{l}\text{Maximize } V(q)\\ \text{subject to } G(X(q)) \leqq 0\end{array}$$

where G represents the aggregate production constraint. This problem gave rise to the first-order conditions ((19) and (22)) which were equivalently stated as

$$\text{(34)}\qquad \begin{aligned}\frac{\partial V}{\partial q_k} &= \lambda \sum p_i \frac{\partial X_i}{\partial q_k}\\ &= -\lambda \frac{\partial}{\partial t_k}\left(\sum t_i X_i\right)\end{aligned} \qquad (k = 1, 2, \ldots, n)$$

Equations (34) were derived only for $k=2, \ldots, n$. But we can see that they hold also for $k=1$; for, on multiplying by q_k and adding, we have

$$\sum_{k=1}^{n}\left[\frac{\partial V}{\partial q_k} - \lambda \sum_i p_i \frac{\partial X_i}{\partial q_k}\right] q_k = 0$$

by the homogeneity of degree 0 of V and the X_i. Equation (34) states that the impact of a price rise on social welfare is proportional to the cost of meeting the change

* Massachusetts Institute of Technology and Nuffield College, respectively. The remainder of the matching footnote in Part I is appropriate here too.

in demand induced by the price rise. Alternatively the impact of a tax increase on social welfare is proportional to the induced change in tax revenue (all calculated at fixed producer prices).

VI. Optimal Tax Structure—One-Consumer Economy

For one consumer and an individualistic welfare function (so that V coincides with v, the indirect utility function of the only consumer in the economy), we can express directly the derivative of social welfare with respect to q_k ($v_k = -\alpha x_k$ where α is the marginal utility of income—see equation (5) of Part I). For this case we can then explore the structure of taxation in more detail. The formulation of the first-order conditions using compensated demand derivatives is due to Paul Samuelson (1951). We begin by stating the familiar Slutsky equation:

$$\frac{\partial x_i}{\partial q_k} = s_{ik} - x_k \frac{\partial x_i}{\partial I} \tag{35}$$

where s_{ik} is the derivative of the compensated demand curve for i with respect to q_k, and $\partial x_i/\partial I$ is the derivative of the uncompensated demand with respect to income (evaluated at $I=0$ in our case). We shall make use of the well-known result that $s_{ik} = s_{ki}$.

Substituting into the first-order conditions (34) we have:

$$\begin{aligned} -\alpha x_k &= -\lambda \frac{\partial}{\partial t_k}\left(\sum t_i x_i\right) \\ &= -\lambda\left(x_k + \sum t_i \frac{\partial x_i}{\partial t_k}\right) \\ &= -\lambda x_k - \lambda \sum t_i s_{ik} \\ &\quad + \lambda x_k \sum t_i \frac{\partial x_i}{\partial I} \\ & \qquad k = 1, 2, \ldots, n \end{aligned} \tag{36}$$

Rearranging terms, we can write this in the form:

$$\frac{\sum_i t_i s_{ik}}{x_k} = \frac{\alpha + \lambda - \lambda \sum t_i \frac{\partial x_i}{\partial I}}{\lambda} \tag{37}$$

The point to be noticed is that the right-hand side of this equation is independent of k. Call it $-\theta$. Finally, using the symmetry of the Slutsky matrix, we write the first-order conditions as:

$$\frac{\sum_i s_{ki} t_i}{x_k} = -\theta \tag{38}$$

Multiplying by $t_k x_k$ and summing, we obtain

$$\theta \sum_k t_k x_k = -\sum_{k,i} t_k s_{ki} t_i \geq 0, \tag{39}$$

by the negative semi-definiteness of the Slutsky matrix. Thus θ has the same sign as net government revenue.

The left-hand side of (38) is the percentage change in the demand for good k that would result from the tax change if producer prices were constant, the consumer were compensated so as to stay on the same indifference curve, and the derivatives of the compensated demand curves were constant at the same level as at the optimum point:

$$\begin{aligned} \Delta x_k &= \sum_i \int_0^{t_i} \frac{\partial x_k}{\partial t_i} dt_i = \sum_i \int_0^{t_i} s_{ki} dt_i \\ &= \sum_i s_{ki} \int_0^{t_i} dt_i = \sum_i s_{ki} t_i \end{aligned} \tag{40}$$

In fact, it is not possible for all these derivatives to be constant. But if the optima taxes are small, it is approximately true that the optimal tax structure implies an equal percentage change in compensated demand at constant producer prices.

We can also calculate the actual change in demand arising from the tax structure (assuming price derivatives of demand and production prices are constant) by resubstituting from the Slutsky equation (35) Then, upon substitution, we have:

$$\sum_i \frac{\partial x_k}{\partial q_i} t_i + \frac{\partial x_k}{\partial I} \sum t_i x_i = -\theta x_k;$$

or

$$\frac{\sum_i \frac{\partial x_k}{\partial q_i} t_i}{x_k} = -\theta - x_k^{-1} \frac{\partial x_k}{\partial I} \sum t_i x_i \tag{41}$$

The actual changes in demand (again assuming constant derivatives) induced by the tax structure differ from proportionality with a larger than average percentage fall in demand for goods with a large income derivative.

Three-Good Economy

In the case of a three-good economy, we can obtain an expression for the relative ad valorem tax rates of the two taxed goods. This argument is similar to that of W. J. Corlett and D. C. Hague, who discussed the direction of movement away from proportional taxation that would increase utility. In the three-good case, with good one untaxed, the first-order conditions (38) become

$$\begin{aligned} s_{22}t_2 + s_{23}t_3 &= -\theta x_2 \\ s_{32}t_2 + s_{33}t_3 &= -\theta x_3 \end{aligned} \tag{42}$$

Solving these equations we have

$$t_2 = \theta \frac{s_{23}x_3 - s_{33}x_2}{s_{22}s_{33} - s_{23}^2}, \quad t_3 = \theta \frac{s_{32}x_2 - s_{22}x_3}{s_{22}s_{33} - s_{23}^2} \tag{43}$$

Notice that the denominator here is positive, by the properties of the Slutsky matrix. We convert these into elasticity expressions, defining the elasticity of compensated demand by

$$\sigma_{ij} = \frac{q_j s_{ij}}{x_i} \tag{44}$$

Equation (43) can then be written

$$\frac{t_2}{q_2} = \theta'(\sigma_{23} - \sigma_{33}), \quad \frac{t_3}{q_3} = \theta'(\sigma_{32} - \sigma_{22}), \tag{45}$$

where

$$\theta' = \frac{\theta x_2 x_3}{q_2 q_3 (s_{22}s_{33} - s_{23}^2)}$$

We now substitute for σ_{23} and σ_{33}, using the adding-up properties of compensated elasticities,

$$\begin{aligned} \sigma_{23} &= -\sigma_{22} - \sigma_{21}, \\ \sigma_{32} &= -\sigma_{33} - \sigma_{31} \end{aligned} \tag{46}$$

This gives us

$$\begin{aligned} \frac{t_2}{q_2} &= \theta'(\sigma_{21} + \sigma_{22} + \sigma_{33}), \\ \frac{t_3}{q_3} &= \theta'(\sigma_{31} + \sigma_{22} + \sigma_{33}) \end{aligned} \tag{47}$$

The interesting case to consider is where labor ($x_1 < 0$) is the untaxed good, while goods 2 and 3 are consumer goods ($x_2 > 0$, $x_3 > 0$). Then θ' has the same sign as net government revenue. For definiteness, suppose that government revenue is positive so that $\theta' > 0$. Equation (47) shows that

$$\frac{t_2}{q_2} \gtreqless \frac{t_3}{q_3} \text{ according as } \sigma_{21} \lesseqgtr \sigma_{31} \tag{48}$$

The tax rate is proportionally greater for the good with the smaller cross-elasticity of compensated demand with the price of labor. (It is possible that one commodity is subsidized, but it has to be the one with the greater cross-elasticity.)

Examples

The implications of the above model are very diverse, depending upon the nature of the demand functions. A simple example will show how the theory can be used. If we define ordinary demand elasticities by the usual formula

$$\epsilon_{ik} = q_k x_i^{-1} \frac{\partial x_i}{\partial q_k}, \tag{49}$$

we can rewrite the optimal taxation formula in the form

$$v_k = q_k^{-1} \lambda \sum p_i x_i \epsilon_{ik} \tag{50}$$

When the welfare function is individualistic, equation (5) applies, so that equation (50) may be written as

$$(51) \qquad -\alpha q_k x_k = \lambda \sum p_i x_i \epsilon_{ik}$$

or

$$q_k p_k^{-1} = -\frac{\lambda}{\alpha} \sum_i \frac{p_i x_i}{p_k x_k} \epsilon_{ik}$$

If we have a good whose price does not affect other demands (implying a unitary own price elasticity), equation (51) simplifies to yield the optimal tax of that good:

$$(52) \qquad \begin{array}{l} \text{If } \epsilon_{ik} = 0 \ (i \neq k) \quad \text{and} \quad \epsilon_{kk} = -1, \\ \text{then } q_k p_k^{-1} = \lambda \alpha^{-1} \end{array}$$

where $q_k p_k^{-1}$ equals one plus the percentage tax rate. Recalling that α is the marginal utility of income while λ reflects the change in welfare from allowing a government deficit financed from some outside source, their ratio gives a marginal cost (in terms of the numeraire good) of raising revenue. Thus the optimal tax rate on such a good gives the cost to society of raising the marginal dollar of tax.

An example of a utility function exhibiting such demand curves is the Cobb-Douglas, where only labor is supplied. As an example consider:

$$(53) \qquad u(x) = b_1 \log (x_1 + \omega_1) + \sum_{i=2}^{n} b_i \log x_i$$

If we choose labor as the untaxed numeraire, all other goods satisfy (52) and we see that the optimal tax structure is a proportional tax structure.

It is easy to exhibit examples where the optimal tax structure is not proportional. Consider the example:

$$(54) \qquad \begin{array}{l} u(x) = \sum b_i \log (x_i + \omega_i), \\ \sum b_i = 1, \ \omega_i \neq 0 \end{array}$$

The demands arising from these preferences are:

$$(55) \qquad x_i = q_i^{-1} b_i \sum q_j \omega_j - \omega_i$$

Therefore the demand elasticities are:

$$(56) \qquad \begin{array}{ll} \epsilon_{ik} = b_i \omega_k x_i^{-1} \dfrac{q_k}{q_i} & (k \neq i) \\ \epsilon_{kk} = -b_k x_k^{-1} \displaystyle\sum_{j \neq k} \omega_j \frac{q_j}{q_k} & \end{array}$$

Substituting in the formula for the optimal taxes,

$$(57) \qquad \begin{aligned} -\alpha q_k x_k = \\ \lambda\left[\sum_{j \neq k} b_j \frac{p_j}{q_j} \omega_k q_k - b_k \frac{p_k}{q_k} \sum_{j \neq k} \omega_j q_j\right] \\ = \lambda \sum_j \left[b_j \omega_k \frac{p_j q_k}{q_j} - b_k \omega_j \frac{p_k q_j}{q_k}\right] \end{aligned}$$

Since the assumption $\sum b_j = 1$ allows us to write the demand functions (55) in the form:

$$(58) \qquad q_k x_k = \sum_j \left[b_k \omega_j q_j - b_j \omega_k q_k\right],$$

we can deduce from (57) and (58) that

$$(59) \qquad \begin{aligned} \sum_j \bigg[b_j \omega_k q_k \left(\frac{p_j}{q_j} - \frac{\alpha}{\lambda}\right) \\ - b_k \omega_j q_j \left(\frac{p_k}{q_k} - \frac{\alpha}{\lambda}\right)\bigg] = 0 \end{aligned}$$

These equations allow us to calculate p for any given q, and in that way give the optimal taxation rules. In general, taxes will not be proportional. As one example of this, consider the following three-good case.

Sample Calculation

Let us combine the above two examples by considering a three-good economy (one-consumer good and two types of labor) with preferences as in (54). This example will be used to show that limited tax possibilities (represented by the same proportional tax on goods 2 and 3) intro-

duces the desirability of aggregate production inefficiency.

Example e. Assume that preferences satisfy

$$u = \log x_1 + \log (x_2 + 1) + \log (x_3 + 2) \tag{60a}$$

$$x_1 > 0, \quad x_2 > 1, \quad x_3 > -2;$$

while private production possibilities are

$$y_1 + y_2 + y_3 \leqq 0, \tag{60b}$$

$$y_1 \geqq 0, \quad y_2 \leqq 0, \quad y_3 \leqq 0;$$

and the government constraint is

$$1.02z_1 + z_2 \leqq 0 \tag{60c}$$

$$z_1 \geqq 0, \quad z_2 \leqq 0, \quad z_3 \leqq -0.1$$

Thus the government needs good 3 for public use and can produce good 1 from good 2, but only less efficiently than the private sector can.

Since we know that production efficiency is desired, we have

$$q_1 = p_1 = p_2 = 1, \qquad z_1 = z_2 = 0$$

From the first-order conditions (59) and market clearance given the demands (58), we obtain two equations to determine q_2 and q_3:

$$q_2(q_3^{-1} - 1) = 2q_3(q_2^{-1} - 1)$$

$$(q_2 + 2q_3)(q_2^{-1} + q_3^{-1} + 1) = 8.7$$

These have a unique positive solution

$$q_2 = 0.94494, \quad q_3 = 0.90008$$

which give

$$x_1 = 0.9150, \quad x_2 = -0.0316, \quad x_3 = -0.9834$$

$$u = -0.1045$$

If we now require the same tax rate on goods 2 and 3 and at the same time impose production efficiency, then $q_2 = q_3 = q$, and the tax rate is determined by the market clearance equation. We obtain

$$3q + 6 = 8.7; \qquad \text{i.e., } q = 0.9$$

Then demands are

$$x_1 = 0.9, \quad x_2 = 0, \quad x_3 = -1$$

and

$$u = -0.1054$$

Notice that the economy is still on the production frontier even though both input prices are lower in this case. If we introduce inefficiency with $p_2 > 1$, so that $y_2 = 0$ and $x_2 = z_2$, we can increase utility. Market clearance now requires

$$(q_2 + 2q_3)((1.02)^{-1}q_2^{-1} + q_3^{-1} + 1) = 8.7$$

At prices $q_2 = .92$, $q_3 = .90008$ for example, we have, $x_1 = 0.9067$, $x_2 = -0.0144$, $x_3 = -0.9926$, and $u = -0.1051$.

VII. Optimal Tax Structure—Many-Consumer Economy

As we noted in Section III of Part I, the equations for optimal taxation with a single consumer which do not reflect the particular form of V are also valid for many consumers. To pursue the analysis further, we must find an expression for V_k, the derivative of social welfare with respect to the kth consumer price.

With an individualistic welfare function, we have

$$V(q) = W(v^1(q), v^2(q), \ldots, v^H(q)) \tag{61}$$

Differentiating with respect to q_k, we obtain

$$V_k = \sum_h \frac{\partial W}{\partial u^h} v_k^h = -\sum_h \frac{\partial W}{\partial u^h} \alpha^h x_k^h \tag{62}$$

The term α^h is the marginal utility of income of consumer h. Therefore

$$\beta^h = \frac{\partial W}{\partial u^h} \alpha^h \tag{63}$$

is the increase in social welfare from a unit increase in the income of consumer h. We have

$$-V_k = \sum_h \beta^h x_k^h, \tag{64}$$

or the derivative of welfare with respect to a price equals the "welfare-weighted" net consumer demand for commodity k. The necessary condition for optimal taxation makes V_k proportional to the marginal contribution to tax revenue from raising the tax on good k.

$$\sum_h \beta^h x_k^h = \lambda \frac{\partial T}{\partial t_k}, \tag{65}$$

where $T = \sum t_i X_i$ is total tax revenue, and the derivative is evaluated at constant producer prices (i.e., on the basis of consumer excess demand functions alone). We also have the alternative formula

$$\sum_h \beta^h x_k^h = -\lambda \sum_i p_i \frac{\partial X_i}{\partial q_k} \tag{66}$$

Example f. Before turning to interpretations of the optimal tax formulae like those above, let us consider an example.

We will assume that each consumer has a Cobb-Douglas utility function,

$$u^h = b_1^h \log (x_1^h + \omega^h) + \sum_2^n b_i^h \log x_i^h, \qquad \sum_1^n b_i^h = 1 \tag{67}$$

Choosing good 1 as numeraire, we saw in Section VI that with a one-consumer economy, taxation would be proportional. This will not, in general, be true in a many-consumer economy where each consumer has this utility function. The individual demand curves arising from this utility function are:

$$\begin{aligned} x_i^h &= q_i^{-1} b_i^h q_1 \omega^h, \qquad i = 2, 3, \ldots, n \\ x_1^h &= -(1 - b_1^h)\omega^h \end{aligned} \tag{68}$$

Notice that $\partial x_i^h/\partial q_k = 0 \; (k \neq i \neq 1)$ and $\partial x_i^h/\partial q_i = -x_i^h/q_i \; (i \neq 1)$.

Assuming an individualistic welfare function, the first-order conditions (66) are in this case

$$\sum_h \beta^h x_k^h = \lambda p_k q_k^{-1} \sum_h x_k^h \qquad (k = 2, \ldots, n) \tag{69}$$

This implies the following formula:

$$\frac{q_k}{p_k} = \lambda \frac{\sum_h x_k^h}{\sum_h \beta^h x_k^h} = \lambda \frac{\sum_h b_k^h \omega^h}{\sum_h \beta^h b_k^h \omega^h} \qquad (k = 2, \ldots, n) \tag{70}$$

To complete the determination of the optimal taxes, we must find the relationship between λ, p_1, and q_1. This is obtained from the Walras identity. The value of net consumer demand in producer prices is equal to minus the profit in production. (Alternatively, we could determine λ so that the government budget is balanced.) That is

$$-p_1 \sum_h (1 - b_1^h)\omega^h + \sum_{i=2}^n \sum_h p_i q_i^{-1} b_i^h q_1 \omega^h = \gamma, \tag{71}$$

where γ is the maximized profit of production net of government needs $(= \sum_{i=1}^n p_i z_i)$. Substituting from (70) and rearranging, we obtain

$$\begin{aligned} \frac{q_1}{p_1} &= \lambda \frac{\sum_h (1 - b_1^h)\omega^h + \gamma p_1^{-1}}{\sum_{i=2}^n \sum_h \beta^h b_i^h \omega^h} \\ &= \lambda \frac{\sum (1 - b_1^h)\omega^h + \gamma p_1^{-1}}{\sum \beta^h (1 - b_1^h)\omega^h} \end{aligned} \tag{72}$$

The number γp^{-1} is determined by the technology and the government expenditure decision, and therefore depends on p (unless $\gamma = 0$).

Equations (70) and (72) determine the optimal tax rates. If the social marginal utilities, β^h, are independent of taxation, the optimal tax rates can be read off at

once. This is true if W has the special form $\sum_h v^h$; for in that case $\beta^h = 1/\omega^h$. It should be noticed that, although each household's social marginal utility of income is unaffected by taxation, it is desirable to have taxation in general. If households with relatively low social marginal utility of income predominate among the purchasers of a commodity, that commodity should be relatively highly taxed. Although such taxation does nothing to bring social marginal utilities of income closer together, it does increase total welfare.

In general, taxation does affect social marginal utilities of income. The β^h depend on the tax rates, and equations (70) do not, therefore, give explicit formulae for the optimum taxes. In the case $W = -\mu^{-1} \sum_h e^{-\mu v^h}$, $\mu > 0$, so that there is a stronger bias toward equality than in the additive case, it can be verified quite easily that the optimum taxes have to satisfy

$$(73) \quad \frac{q_k}{p_k} \sum_h b_k^h (\omega^h)^{-\mu} \prod_{i=2}^{n} (b_i^h)^{-\mu b_i^h} q_i^{\mu b_i^h} = \lambda \sum_h b_k^h \omega^h \qquad (k = 2, 3, \ldots, n)$$

In this case, marginal utilities of income are brought closer together.[1] It is not immediately obvious from the equations (10) that the q are determined given the p. However, it can be shown that, in the present example, the first-order conditions must have a unique solution.[2] In fact, the relations (70) (along with (72)) would, if followed by government, certainly lead to maximum welfare if production were perfectly competitive, since any state of the economy satisfying these conditions maximizes welfare, and the maximum is unique for the welfare function considered. Unfortunately this convenient property is not general.

From equation (70) we can identify two cases where optimal taxation is proportional. If the social marginal utility of income is the same for everyone ($\beta^h = \beta$, for all h), then equation (70) reduces to $q_k p_k^{-1} = \lambda/\beta$. In this case there is no welfare gain to be achieved by redistributing income, and so no need to tax differently (on average) the expenditures of different individuals. Thus the optimal tax formula has the same form as in the one-consumer case. When the β^h do differ, taxes are greater on commodities purchased more heavily by individuals with a low social marginal utility of income. If, for example, the welfare function treats all individuals symmetrically and if there is diminishing social marginal utility with income, then there is greater taxation on goods purchased more heavily by the rich.

The second case leading to proportional taxation occurs when demand vectors are proportional for all individuals, $x^h = \rho^h x$, and thus $b_k^h = b_k$ for all h. With all individuals demanding goods in the same proportions, it is impossible to redistribute income by commodity taxation implying that the tax structure again assumes the form it has in a one-consumer economy.

Optimal Tax Formulae

The description in Section VI of some possible interpretations of the optimal tax formula carries over to the many-consumer case. Thus, as was true there con-

---

[1] If $\mu < 0$, utilities and marginal utilities are moved further apart.

[2] It is easily verified that $v^h = \delta_h + \sum_i b_i \log (q_1/q_i)$, where the δ_h are constants. Consequently

$$V(q) = -\mu^{-1} \sum_h e^{-\mu\delta_h} \prod_i (q_1/q_i)^{-\mu b_i^h}$$

which is a concave function of $(q_1/q_2, q_1/q_2, \ldots, q_1/q_n)$. Also, aggregate demand is

$$X_i(q) = \sum_h b_i^h \omega^h \cdot (q_1/q_i), \; X_1(q) = -\sum_h (1 - \alpha_i^h)\omega^h$$

If the production set is convex, the set of $(q_1/q_2, \ldots, q_1/q_n)$ for which $(X_1, X_2, \ldots, X_n)$ is feasible is also convex. Thus the optimum q is obtained by maximizing a concave function of $(q_1/q_2, \ldots, q_1/q_n)$ over a convex set, and is therefore uniquely defined by the first-order conditions.

sumer price elasticities but not producer price elasticities enter the equations, and at the optimum the social marginal utility of a price change is proportional to the marginal change in tax revenue from raising that tax, calculated at constant producer prices. Analysis of the change in demand can also be carried out, but is naturally more complicated. Assuming an individualistic welfare function, the first-order conditions can be written[3]

$$\sum \beta^h x_k^h = \lambda \sum_h \sum_i t_i \frac{\partial x_i^h}{\partial q_k} + \lambda \sum_h x_k^h \tag{74}$$

From the Slutsky equation, we know that

$$\begin{aligned} \frac{\partial x_i}{\partial q_k} &= s_{ik} - x_k \frac{\partial x_i}{\partial I} = s_{ki} - x_k \frac{\partial x_i}{\partial I} \\ &= \frac{\partial x_k}{\partial q_i} - x_k \frac{\partial x_i}{\partial I} + x_i \frac{\partial x_k}{\partial I} \end{aligned} \tag{75}$$

Substituting from (75) in (74) we can write the optimal tax formula as equation (76). Rearranging terms we can write equation (76) as (77). With constant producer prices, equation (77) gives the change in demand as a result of taxation for a good with constant price-derivatives of the demand function (or for small taxes). Considering two such goods, we see that the percentage decrease in demand is greater for the good the demand for which is concentrated among:

(1) individuals with low social marginal utility of income,
(2) individuals with small decreases in taxes paid with a decrease in income,
(3) individuals for whom the product of the income derivative of demand for good k and taxes paid are large.

[3] We neglect the possibility of a free good when the first-order condition would be an inequality.

VIII. Other Taxes

Thus far we have examined the combined use of public production and commodity taxation as control variables. It is natural to reexamine the analysis when additional tax variables are included in those controlled by the government. In particular, in the next subsection we will briefly consider income taxation; but first, let us examine a general class of taxes such that the consumer budget constraint depends on consumer prices and on tax variables. We shall replace the budget constraint $\sum q_i x_i = 0$ by the more general constraint $\phi(x, q, \zeta) = 0$, where ζ represents a shift parameter to reflect the choice among different systems of additional taxation (for example, the degree of progression in the income tax). Let us note that this formulation continues to assume that all taxes are levied on consumers and that there are no profits in the economy.

The key assumption to permit an extension of the analysis above is an independence of the two constraints on the planner. We need to assume that the choice of tax variables does not affect the production

$$\sum_h \beta^h x_k^h = \lambda \sum_h \sum_i t_i \frac{\partial x_k^h}{\partial q_i} + \lambda \sum_h \sum_i t_i \left(x_i^h \frac{\partial x_k^h}{\partial I} - x_k^h \frac{\partial x_i^h}{\partial I} \right) + \lambda \sum_h x_k^h \tag{76}$$

$$\frac{\sum_h \sum_i t_i \frac{\partial x_i^h}{\partial q_i}}{\sum_h x_k^h} = \frac{1}{\lambda} \frac{\sum_h \beta^h x_k^h}{\sum_h x_k^h} - 1 + \frac{\sum_h \left(\sum_i t_i \frac{\partial x_i^h}{\partial I} \right) x_k^h}{\sum_h x_k^h} - \frac{\sum_h \left(\sum_i t_i x_i^h \right) \frac{\partial x_k^h}{\partial I}}{\sum_h x_k^h} \tag{77}$$

possibilities, and further that the choice of a production point does not affect the set of possible demand configurations. In particular, this formulation implies that producer prices do not affect consumer budget constraints. Thus the income tax, to fit this formulation, needs to be levied on the wages that consumers receive, not on the cost of wages to the firm. Similarly it is assumed that there are no sales tax deductions from the income tax base.

We know already that in such a case, optimal production is efficient. We may therefore concentrate upon the case in which all production is controlled by the government, and the production constraint is that $x_1 = g(x_2, x_3, \ldots, x_n)$. We have to choose $q_2, q_3, \ldots, q_n, \zeta$ to

$$\text{(78)} \quad \text{maximize } V(q, \zeta) \text{ subject to } X_1(q, \zeta) = g(X_2(q, \zeta), \ldots, X_n(q, \zeta))$$

As before we introduce a Lagrange multiplier λ. Differentiation with respect to q_k yields the familiar

$$\text{(79)} \qquad V_k = \lambda \sum_i p_i \frac{\partial X_i}{\partial q_k},$$

where the producer price p_i is $\partial g/\partial x_i$ $(i = 2, 3, \ldots, n)$, and $p_1 = 1$. Differentiation with respect to the new tax variable provides the similar equation

$$\text{(80)} \qquad \frac{\partial V}{\partial \zeta} = \lambda \sum_i p_i \frac{\partial X_i}{\partial \zeta}$$

We have an alternative form for (79), namely,

$$\text{(81)} \qquad V_k = -\lambda \frac{\partial T}{\partial t_k}$$

In exactly the same way, we obtain from (80) a formula involving the effect of the new tax on total tax revenue,

$$\text{(82)} \qquad V_\zeta = -\lambda \frac{\partial T}{\partial \zeta}$$

Income Taxation

Nothing that we have said suggests that commodity taxation is superior to income taxation. The analysis has only considered the best use of commodity taxation. It is natural to go on to ask how one employs both commodity taxation and income taxation. The formulation of income taxation raises a problem. If the planners are free to select any income tax structure and if there are a finite number of tax payers, the tax structure can be selected so that the marginal tax rate is zero for each taxpayer at his equilibrium income (although this does not necessarily bring the economy to the full welfare maximum). This eliminates much of our problem, but like lump sum taxation, seems to be beyond the policy tools available in a large economy. The natural formulation of this problem is for a continuum of tax payers, since then no man can have a tax schedule tailor-made for him. (This approach is taken by Mirrlees.) However, we shall here take the alternative route by assuming a limited set of alternatives for the income tax structure.

If only commodity taxation is possible, the tax paid by a household that purchases a vector x^h is

$$\text{(83)} \qquad T^h = \sum_i t_i x_i^h$$

To add income taxation to the tax structure, we can select a subset of commodities, L, e.g., labor services, and tax the value of transactions on this subset, so that

$$I^h = \sum_{i \text{ in } L} q_i x_i^h$$

where I is "taxable income." Then

$$\text{(84)} \qquad T^h = \sum_i t_i x_i^h + \tau(I^h, \zeta),$$

where τ is a fixed continuously differentiable function depending on a parameter ζ, and is the same for all consumers. With a

tax on services (x_i negative) we would expect τ to be decreasing in its tax base, with a derivative between zero and minus one. In terms of the notation employed above, we can define the budget constraint $\phi(x^h, q, \zeta)$ by

$$\begin{aligned}\phi(x^h, q, \zeta)& \\ = &\sum p_i x_i^h + T^h \\ = &\sum q_i x_i^h + \tau\left(\sum_{i \text{ in } L} q_i x_i^h, \zeta\right)\end{aligned} \tag{85}$$

Here we can regard q and ζ as the policy variables. Thus the consumer's budget constraint can be expressed in a form depending on consumer prices and independent of producer prices.

The first-order conditions for optimal income taxation are just the conditions (79) and (80), interpreted for this special case. The social marginal utility of a tax variable change is proportional to the marginal change in tax revenue calculated at constant producer prices. In the case of an individualistic welfare function, we can give more explicit formulae for the welfare derivatives, V_k and V_ζ:

$$V_k = \sum_h \beta^h x_k^h \left(1 + \delta_k \frac{\partial \tau^h}{\partial I}\right) \tag{86}$$

$$V_\zeta = \sum_h \beta^h \frac{\partial \tau^h}{\partial \zeta}, \tag{87}$$

where $\delta_k = 1$ if k is in L, 0 if k is not in L; and $\tau^h = \tau(I^h, \zeta)$.

These equations are derived from the first-order conditions for maximizing u^h subject to $\phi = 0$, noticing that, for example, the budget constraint implies that

$$\sum_k \frac{\partial \phi}{\partial x_k} \frac{\partial x_k}{\partial \zeta} + \frac{\partial \phi}{\partial \zeta} = 0$$

Combining (82) and (87), we obtain

$$\sum \beta^h \frac{\partial \tau^h}{\partial \zeta} = \lambda \frac{\partial T}{\partial \zeta} \tag{88}$$

Thus, at the optimum, for any two different kinds of change in the income tax structure, the social-marginal-utility weighted changes in taxation (consumer behavior held constant) are proportional to the changes in total tax revenue (both income and commodity tax revenue, calculated at fixed producer prices, with consumer behavior responding to the price change).

IX. Public Consumption

From the start, we have considered the government production decision as constrained by $G(z) \leq 0$. The presence of a fixed bundle of public consumption was therefore included in the model (and would show itself by $G(0)$ being positive). This is unsatisfactory and was assumed to keep as uncluttered as possible a naturally complicated problem. We can now consider a choice among vectors of public consumption which affect social welfare directly. (We shall assume that the government controls all production, thus ignoring public expenditures which affect private production rather than consumer utility.) Let us denote by e the vector of public consumption expenditures. (Items of public consumption which are difficult to measure can be described by the inputs into their production.) The presence of public consumption alters our problem in three ways. First, public consumption represents public production (or purchases) which are not supplied to the market. Thus market clearance becomes $X = z - e$.

Second, the presence of public consumption affects private net demand, which must now be written $X(q, e)$. Third, the level of public consumption directly affects the social welfare function (by affecting individual utility in the case of an individualistic welfare function).

We can restate the basic maximization problem as

(89) $\underset{q,e}{\text{Maximize}}\ V(q, e)$

subject to $G(X(q, e) + e) \leqq 0$

The presence of e in the problem will not affect the equations obtained by differentiating a Lagrangian expression with respect to q. Thus the presence of alternative bundles of public consumption does not alter the rules for the optimal tax structure. Nor would we expect it to affect the conditions which imply production efficiency at the optimum. We can therefore replace the inequality in (89) with an equality and differentiate the Lagrangian expression with respect to e_k:

$$\frac{\partial V}{\partial e_k} - \lambda\left[\sum G_i \frac{\partial X_i}{\partial e_k} + G_k\right] = 0 \tag{90}$$

Since

$$\begin{aligned} \sum G_i \frac{\partial X_i}{\partial e_k} &= \sum p_i \frac{\partial X_i}{\partial e_k} = \sum (q_i - t_i) \frac{\partial X_i}{\partial e_k} \\ &= \frac{\partial}{\partial e_k}\left(\sum q_i X_i - \sum t_i X_i\right) \\ &= -\frac{\partial}{\partial e_k}\left(\sum t_i X_i\right), \end{aligned} \tag{91}$$

we can write (90) as

$$\frac{\partial V}{\partial e_k} = -\lambda \frac{\partial}{\partial e_k}\left(\sum t_i X_i\right) + \lambda G_k \tag{92}$$

Equations (92) show how the optimal level of public consumption depends on:

(i) the direct contribution of public consumption to welfare (measured by $\partial V/\partial e_k$);

(ii) the effect of public consumption on tax revenue (measured by $\partial \sum t_i X_i/\partial e_k$); and

(iii) the direct cost of public consumption (G_k).

There are three differences between this theory and that of public goods in the presence of lump sum taxation (as developed, for example, by Samuelson (1954)). Because social marginal utilities of income are not equated, the expression $\partial V/\partial e_k$ cannot be reduced to a sum of marginal rates of substitution, but depends on the weights given to the different beneficiaries of public consumption:

$$\frac{\partial V}{\partial e_k} = \sum_h \frac{\partial W}{\partial u^h} \frac{\partial u^h}{\partial e_k} \tag{93}$$

Second, the cost associated with the raising of government revenue implies that the impact of public consumption on revenue is a relevant part of the first-order conditions. Third, for the same reason, the cost of public consumption is measured in terms of the cost to the government of raising revenue to finance the expenditures (in terms of the one-consumer equation, λ may not be equal to α, the marginal utility of income).

The first-order conditions for the provision of public goods can be expressed in another way, showing the relationships between the marginal cost and "willingness to pay." Write r_k^h for the marginal rate of substitution between public good k and income for the hth household. Then $\partial u^h/\partial e_k = \alpha^h r_k^h$, where α^h is the hth household's marginal utility of income. The social marginal utility of the hth household's income, β^h, is $(\partial W/\partial u^h)\alpha^h$. Consequently, from (93)

$$\frac{\partial V}{\partial e_k} = \sum_h \beta^h r_k^h \tag{94}$$

Then, from (92)

$$G_k = \sum_h \left[\frac{\beta^h}{\lambda} r_k^h + \frac{\partial}{\partial e_k} \sum_i t_i x_i^h\right] \tag{95}$$

Thus the marginal cost of producing the public good should be equated to a sum, over all households, of the price which the household is just willing to pay for a

marginal increment in the level of provision, weighted by the marginal "social worth" of the household's income, and adjusted for the effect of the level of provision on net tax payments by the household.[4]

In the discussion of public consumption thus far it has been assumed that there were no possible fees associated with the provision of public goods. This would be appropriate for national defense or preventive medicine, but not for goods where licenses can be required from users. The optimal level of license fees will not, in general, be zero. Indeed we may be able to associate with any good more complicated pricing mechanisms than the single fixed price considered above. In particular, there are the familiar examples of two-part tariffs (a license fee for use of a facility plus a per unit charge on the amount of use), and prices depending on quantity of sales. Formally these can be treated in a fashion similar to the income taxes considered above; the set of goods over which the tax is defined is now a consumption good rather than labor. With a two-part tariff, this would imply a tax function which was not continuous at the origin.

Presumably the introduction of more general pricing and taxing schemes gives an opportunity for increasing social welfare, just as the progressive income tax gives such an opportunity. In practice, the ignored costs of tax administration may severely limit the number of complicated pricing schemes which can increase welfare. We would expect the analysis done above to be basically unchanged by the addition of these possibilities, although a two-part tariff will cause aggregate demand to have discontinuities. In practice we would expect these discontinuities to be small relative to aggregate demand, and formally, they could be eliminated by the device of a continuum of consumers.

X. The Optimal Taxation Theorem

In the earlier discussion, we employed calculus techniques to obtain the first-order conditions for the optimal tax structure. However, the valid use of Lagrange multipliers is subject to certain restrictions, which in the present case have no very obvious economic significance. This section provides a rigorous analysis of conditions under which the tax formulae (34) are indeed necessary conditions for optimality, and in particular provides economically meaningful assumptions that ensure their validity. The reader should be warned that the discussion is highly technical.

One might hope to provide a rigorous analysis by using the well-known Kuhn-Tucker theorem for differentiable (not necessarily concave) functions. This theorem requires a certain "constraint qualification" to be satisfied. Let us apply it and see how far we get. We wish to

$$\text{Maximize } V(q)$$
$$\text{subject to } g(X(q)) \leqq 0 \quad \text{and} \quad q \geqq 0,$$

where g is a (vector) production constraint such that $g(X) \leqq 0$ if, and only if, X is in G. Given that V, X, and g are differentiable, and that the Kuhn-Tucker constraint qualification is satisfied, we have the first-order conditions

$$(96) \quad V'(q^*) = \frac{\partial V}{\partial q} \leqq p \cdot \frac{\partial X}{\partial q} = p \cdot X'(q^*),$$

where $p = \lambda \cdot g'(X(q^*))$ for a vector of Lagrange multipliers λ, and is therefore a support or tangent hyperplane to G at $X(q^*)$. Since V and X are homogeneous

[4] Another case can be treated in a similar manner: that of limited government production of a good, which is also being produced privately, when government production is given away rather than being sold. Since the government production rule given above does not reduce to the first-order condition in producer prices, we would not find aggregate production efficiency for the sum of these two sources of production.

of degree zero, $[V'(q^*)-p\cdot X'(q^*)]\cdot q^*=0$: consequently $\partial V/\partial q_i=p\cdot(\partial X/\partial q_i)$ for i such that $q_i^*>0$.

To express the first-order conditions in this form, we naturally expect to assume that V and X are continuously differentiable: to that extent, the differentiability assumptions are innocuous. The assumption that the production set can be described by a finite number of continuously differentiable inequality constraints that satisfy the constraint qualification is less satisfactory. The constraint qualification is an assumption about the functions g: one can violate it by changing the functions g without changing the actual constraint set, G. Some such assumption is required to avoid not unreasonable counter-examples, as we shall see below. But it is not at all obvious how one would check whether a particular example that failed to satisfy the constraint qualification could be put right by describing G by a better behaved set of inequalities. We should like to use a constraint qualification that depends on the properties of the set G (and X) rather than the particular functions g; and we should like the assumption to be more amenable to economic interpretation. The theorem we prove below contains such an assumption, for the case where G is convex and has an interior.

Before stating the theorem let us consider an example in which the first-order conditions are not satisfied at the optimum.

Example g. Consider the one-consumer economy. In the case shown in Figure 10, the offer curve is tangent to the production frontier at the optimum production point. As q varies, the vector $X(q)$ traces out the offer curve. Thus, holding q_2 constant, the vector $\partial X(q)/\partial q_1$ is tangent to the offer curve at $X(q^*)$. Therefore if p is the vector of producer prices, which is tangent to the production frontier at $X(q^*)$, $p\cdot\partial X(q^*)/\partial q_1=0$. The same is true for the derivatives with respect to q_2. But there is no reason why $V'(q^*)$ should be zero: therefore the above first-order conditions may not be satisfied at the optimum.

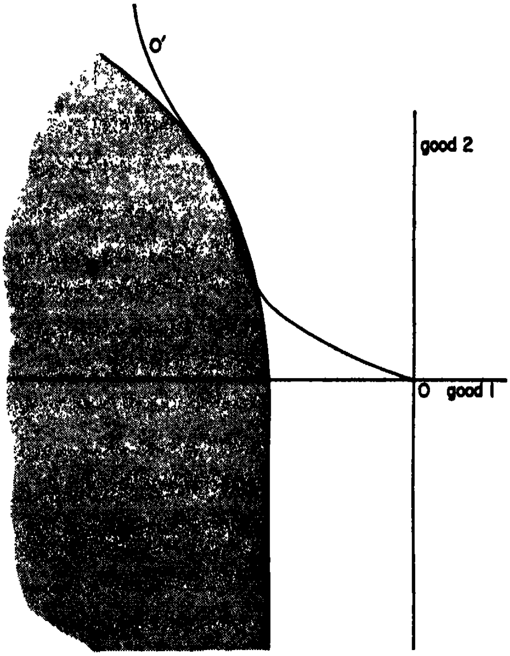


FIGURE 10

We shall make an assumption ruling out tangency between the frontier of the production set and the offer curve:

> For any p, q ($q\geqq 0$, $p\neq 0$) such that $X(q)$ is in G and $p\cdot X(q)\geqq p\cdot x$ for all x in G, $p\cdot X'(q)\geqq 0$.

The qualification takes this particular form because we also have the constraint $q\geqq 0$. Let us note that for $q>0$ the condition $p\cdot X(q)\geqq 0$ is equivalent to $p\cdot X'(q)\neq 0$, because X is homogeneous of degree zero. The qualification asserts that *for any possible competitive equilibrium (under commodity taxation) there is a consumer price change which will decrease the value of equilibrium demand, measured in producer*

prices. By the aggregate consumer budget constraint, $q\cdot X=(p+t)\cdot X=0$. Therefore the assumption says that at any possible equilibrium point on the production frontier, it is possible to increase tax revenue. Thus the first-order conditions may not be applicable if the optimal point represents a local tax revenue maximum. Returning to example g, we see that $p\cdot X'=0$ at the optimum, or equivalently $\partial(t\cdot X)/\partial t=0$, although the derivatives of V are not necessarily zero there.

We now state and prove the theorem.[5]

THEOREM 5: *Assume an optimum, (X^*, q^*) exists; that $V(q)$ and $X(q)$ are continuously differentiable; and that G is convex and has a nonempty interior. Assume furthermore that there is no pair of price vectors (p, q) for which*

$$\begin{aligned} &X(q) \text{ maximizes } p\cdot x \text{ for } x \text{ in } G,\\ (97)\qquad &p \neq 0, \quad \text{and}\\ &p\cdot X'(q) \geqq 0 \end{aligned}$$

Then there exists p^* such that

$$X^* \text{ maximizes } p^*\cdot x \text{ for } x \text{ in } G, \quad \text{and}$$
$$V'(q^*) \leqq p^*\cdot X'(q^*)$$

PROOF:

Let $P=\{p \mid p\cdot X^*\geqq p\cdot x, \text{ all } x \text{ in } G\}$. P is the cone of normals to G at X^*, including the zero vector. It is a nonempty, closed, convex cone.

We write V' for $V'(q^*)$ and X' for $X'(q^*)$. Consider the set

$$B = \{v \mid v \leqq p\cdot X', \text{ some } p \text{ in } P\}$$

We have to show that V' is in B. We do this by showing first, that if V' is in $\bar{B}$, the closure of B, in fact V' is in B; and then that V' must be in $\bar{B}$.

If V' is in $\bar{B}$, there exist sequences $\{v_n\}$ and $\{p_n\}$, p_n in P, such that

$$(98)\qquad \begin{aligned} &v_n \leqq p_n\cdot X',\\ &v_n \to V' \qquad (n\to\infty) \end{aligned}$$

Either $\{p_n\}$ is bounded or it is not. If not, we can find a subsequence on which

$$\|p_n\| \to \infty, \qquad \frac{p_n}{\|p_n\|} \to \bar{p} \neq 0$$

Then, dividing (98) by $\|p_n\|$ and letting $n\to\infty$ on the subsequence, we obtain $\bar{p}\cdot X'\geqq 0$ while $\bar{p}$, $\neq 0$, is in P. This possibility is excluded by assumption (97). Therefore $\{p_n\}$ is bounded, and has a limit point p, in P. Equation (98) implies that $V'\leqq p\cdot X'$. The conclusion of the theorem is thus established on the assumption that V' is in $\bar{B}$.

Suppose, on the contrary, that V' is not in $\bar{B}$. We shall derive a contradiction by a sequence of lemmas.

LEMMA 5.1:

$\bar{B}$ is pointed. That is, v and $-v$ both belong to $\bar{B}$ only if $v=0$.

PROOF:

If v, $-v$ is in $\bar{B}$, we have sequences such that

$$(99)\qquad \overset{1}{v}_n \leqq \overset{1}{p}_n\cdot X', \qquad \overset{2}{v}_n \leqq \overset{2}{p}_n\cdot X',$$

$$(100)\qquad \overset{1}{v}_n \to v, \qquad \overset{2}{v}_n \to -v$$

If $v\neq 0$, it cannot be the case that p_n^1 and p_n^2 both tend to zero. Suppose, for example, p_n^1 does not, and take a subsequence on which

[5] It should be noticed that when the constrained optimum is (locally) an unconstrained maximum, the producer prices satisfying the theorem are zero. This happens if optimal production is in the interior of the production set and may happen if it is on the frontier. The theorem can be weakened in a complicated manner by replacing the nontangency qualification by two conditions. One is an analog of the Kuhn-Tucker Constraint Qualification providing for the existence of an arc in the attainable set. The other use of nontangency occurs when V' is in $\bar{B}$ but not in B. If it is assumed that when there is tangency, the cone of normals is polyhedral, B will be closed. The Kuhn-Tucker theorem is then a special case of the weakened version of theorem 5 when G is the nonnegative orthant. The Kuhn-Tucker theorem is very much easier to prove, however.

$$\|p_n^1\| \to \pi_1 \leqq \infty,$$

$$p_n^1/\|p_n^1\| \to p^1, \neq 0$$

If $p_n^1+p_n^2 \to 0$, $p_n^2/\|p_n^1\| \to -p^1$, and therefore $-p^1$ is in P. This is impossible, since, G having a nonempty interior, P is pointed. (If p, $-p$ are in P, $p \cdot x$ is constant for x in G, but a hyperplane has no interior.) We can therefore take a subsequence on which

$$\|p_n^1 + p_n^2\| \to \pi, \qquad 0 < \pi \leqq \infty,$$

$$\frac{p_n^1 + p_n^2}{\|p_n^1 + p_n^2\|} \to p, \neq 0, \in P$$

From (99) (adding and dividing by $\|p_n^1+p_n^2\|$) and (100), we now have

$$p \cdot X' \geqq \operatorname{Lim} \frac{v_n^1 + v_n^2}{\|p_n^1 + p_n^2\|} = 0 \tag{101}$$

This contradicts (97), since p is in P and $p \neq 0$, and thereby establishes the lemma.

LEMMA 5.2: *If C is a pointed, closed, convex cone, there exists a vector p such that for all non-zero z in C, $p \cdot z < 0$.*

PROOF:

By the duality theorem for convex cones $C^{++} = C$, where C^+ is the dual cone, $\{p | p \cdot z \leqq 0, z \text{ is in } C\}$. Clearly, if C^+ is pointed, C has a nonempty interior: for if interior C is empty, $p \cdot z = 0$ for some nonzero p and all z in C, and then p and $-p$ both belong to C^+. Under the assumptions of the theorem, C is closed and pointed. Therefore C^{++} is pointed, and C^+ has an interior point p.

$$p \cdot z < 0 \quad \text{(all nonzero } z \text{ in } C)$$

Otherwise, if $p \cdot z = 0$, we can easily find a sequence $\{p_n\}$ on which $p_n \to p$ and $p_n \cdot z > 0$, so that p_n is not in C^+.

LEMMA 5.3: *If V' is not in $\bar{B}$, there exists r such that*

$$V' \cdot r > 0 \tag{102}$$

$$v \cdot r < 0 \qquad (v \in B) \tag{103}$$

PROOF:

The closed convex cone $\bar{B} + \{\lambda V' | \lambda \leqq 0\}$ is pointed. Thus there exists an r such that

$$v \cdot r + \lambda V' \cdot r < 0$$

$$(v \in \bar{B}, \lambda \leqq 0, v, \lambda \text{ not both zero})$$

Putting $v=0$ and $\lambda=-1$ we obtain (102); putting $\lambda=0$ we obtain (103).

LEMMA 5.4: *Let r be a vector satisfying (102) and (103). For some $\delta > 0$,*

$$X(q^* + \theta r) \in G \quad (0 \leqq \theta \leqq \delta) \tag{104}$$

PROOF:

Assume not. Then for some sequence $\{\theta_n\}$, $\theta_n > 0$, $\theta_n \to 0$,

$$X(q^* + \theta_n r) \notin G$$

Since G is convex, this implies that

$$X(q^*) + \frac{\lambda}{\theta_n}[X(q^* + \theta_n r) - X(q^*)] \notin G$$

for $\lambda \geqq \theta_n$. Letting $n \to \infty$, we deduce, for any $\lambda > 0$, that

$$X(q^*) + \lambda X' \cdot r = \operatorname{Lim}_{n\to\infty}\left[X(q^*) + \lambda \frac{X(q^* + \theta_n r) - X(q^*)}{\theta_n}\right]$$

is not in the interior of G. It follows that the half-line $\{X(q^*) + \lambda X' \cdot r | \lambda > 0\}$ can be separated from the interior of G by a hyperplane with normal $p \neq 0$:

$$p \cdot X(q^*) + \lambda p \cdot X' \cdot r \geqq p \cdot x$$

$$(\lambda > 0, x \in \operatorname{Int} G)$$

Letting $\lambda \to 0$ we have $p \in P$. Letting $x \to X^*$ we have

$$p \cdot X' \cdot r \geqq 0,$$

which contradicts (103) since $p \cdot X'$ is in B. The lemma is proved.

Since q^* is optimal, (104) implies that

$$V(q^* + \theta r) \leqq V(q^*) \qquad (0 \leqq \theta \leqq \delta)$$

Therefore,

$$V' \cdot r = \lim_{\theta \to 0} \frac{1}{\theta} [V(q^* + \theta r) - V(q^*)]$$
$$\leqq 0$$

This, however, contradicts (102). The hypothesis of Lemma 5.3, that $V' \in \bar{B}$, is therefore false. The proof of the theorem is thus complete.

In reaching our results that the first-order conditions for optimum taxes (96) hold in general, we have assumed that the production set, G, is convex. But one common argument for government control of production is nonconvexity of the production set. This is not a question we are primarily concerned with in this paper. However, some extensions of the theorem do hold. As an example, assume the frontier of G is differentiable at X^*, so that p can be uniquely defined as the normal at X^* and that G is not thin in the neighborhood of X^*—i.e., there exists a ball with center on the normal through X^*, contained in G and containing X^*. Applying the theorem to this ball we get the validity of the first-order conditions (96) using the producer prices defined by the normal.

As in general welfare economics, two uniqueness problems may arise when considering the application of the first-order conditions to achieve an optimum. In the first place, there may be more than one pair of price vectors, (p, q), that satisfy the first-order conditions and allow markets to be cleared. This is similar to the problem that arises when we attempt to define optimum production and distribution by first-order conditions in the presence of a non-convex production set. It is noteworthy that, if lump sum transfers are excluded as a feasible policy, this problem may arise even when the production set is convex. There is no reason why the demand functions should have any of the nice convexity properties which ensure that first-order conditions imply global maximization. Only in particular cases, such as that discussed in footnote 2 above (where rigorous argument is possible without appeal to theorem 5), will the first-order conditions lead to a unique solution.

The second problem is that the tax policies one might like to employ may not uniquely determine the behavior of the system. The lump sum redistribution of wealth required in standard welfare economics does not carry with it any guarantee that the desired competitive equilibrium is the unique one consistent with the optimal wealth distribution (although if the wrong equilibrium is achieved, this should be easily noticed). Similarly, in the present case, if we employ taxes rather than consumer prices as the government control variables, the equilibrium of the economy may not be unique.[6] But if consumer prices are used as the control variables—and why not?—the demand functions give us a unique equilibrium position, so long as preferences are strictly convex.

XI. Concluding Remarks

Welfare economics has usually been concerned with characterizing the best of attainable worlds, accepting only the basic technological constraints. As economists have been aware, the omitted constraints on communication, calculation, and administration of an economy (not to mention political constraints) limit the direct applicability of the implications of this theory to policy problems, although great insight into these problems has certainly been acquired. We have not at-

[6] For a discussion of multiple equilibria in a related problem, see E. Foster and H. Sonnenschein.

tempted to come directly to grips with the problem of incorporating these complications into economic theory. Instead, we have explored the implications of viewing these constraints as limits on the set of policy tools that can be applied. There are many sets of policy tools which might be examined in this way. Specifically, we have assumed that the policy tools available to the government include commodity taxation (and subsidization) to any extent. For these tools we have derived the rules for optimal tax policy and have shown the desirability of aggregate production efficiency, in the presence of optimal taxation. We have also considered expansion of the set of policy tools in such a way that we continue to have the condition that production decisions do not change the class of possible budget constraints. For example, this condition is still preserved when one includes poll taxes, progressive income taxation, regional differences in taxation, taxation on transactions between consumers, and most kinds of rationing. This type of expansion of the set of policy tools does not alter the desirability of production efficiency, nor does it alter the conditions for the optimal commodity tax structure, although in general the tax rates themselves will change. We have, unfortunately, ignored the cost of administering taxes. Presumably optimization by means of sets of policy tools that do not, because the cost of administration, include the full scope of commodity taxation, will not lead to the same conclusions.

Let us briefly consider the type of policy implications that are raised by our analysis. In the context of a planned economy our analysis implies the desirability of using a single price vector in all production decisions, although these prices will, in general, differ from the prices at which commodities are sold to consumers.

As an application of this analysis to a mixed economy, let us briefly examine the discussion of a proper criterion for public investment decisions. As has been widely noted, there are considerable differences in western economies between the intertemporal marginal rates of transformation and substitution. This has been the basis of analyses leading to investment criteria which would imply aggregate production inefficiency because they employ an interest rate for determining the margins of public production which differs from the private marginal rate of transformation. One argument used against these criteria is that the government, recognizing the divergence between rates of transformation and substitution, should use its power to achieve the full Pareto optimum, bringing these rates into equality. When this is done, the single interest rate then existing will be the appropriate rate to use in public investment decisions. We begin by presuming that the government does not have the power to achieve any Pareto optimum that it chooses. Then from the maximization of a social welfare function, we argued that the government will, in general, prefer one of the non-Pareto optima to the Pareto optima, if any, that can be achieved. At the constrained optimum, which is the social welfare function maximizing position of the economy for the available policy tools, we saw that the economy will still be characterized by a divergence between marginal rates of substitution and transformation, not just intertemporally, but also elsewhere, e.g., in the choice between leisure and goods. However, we concluded that in this situation we desired aggregate production efficiency. This implies the use of interest rates for public investment decisions which equate public and private marginal rates of transformation.

We have obtained the first-order conditions for public production, but we have not considered the correct method of evaluating indivisible investments. This

is one problem that deserves examination. In examining the optimal tax structure, we have briefly considered the tax rates implied by particular utility functions. This analysis should be extended to more general and more interesting sets of consumers. Further, we have not examined in any detail the uniqueness and stability of equilibrium, that is, the question whether there are means of achieving in practice an equilibrium which is close to the optimum.

Finally, we would like to emphasize the assumptions which seem to us most seriously to limit the applications of this theory.[7] We have assumed no costs of tax administration and no tax evasion. And we have assumed constant-returns-to-scale and price-taking, profit-maximizing behavior in private production. Pure profits (or losses) associated with the violation of these assumptions imply that private production decisions directly influence social welfare by affecting household incomes. In such a case, it would presumably be desirable to add a profits tax to the set of policy instruments. Nevertheless, aggregate production efficiency would no longer be desirable in general; although it may be possible to get close to the optimum with efficient production if pure profits are small. We hope, nevertheless, that the methods and results of this paper have shown that economic analysis need not depend on the simplifying, but unrealistic, assumption that the perfect capital levy has taken place.[8]

[7] These assumptions are viewed in the context of equilibrium theory. There is no need here to go into the limitations inherent in current equilibrium theory.

[8] A recent paper by Clarence Morrison also deals with marginal cost pricing as a special case of optimal pricing.

REFERENCES

W. J. Corlett and D. C. Hague, "Complementarity and the Excess Burden of Taxation," *Rev. Econ. Stud.*, 1953, *21*, No. 1, 21–30.

E. Foster and H. Sonnenschein, "Price Distortion and Economic Welfare," *Econometrica*, Mar. 1970, *38*, 281–97.

H. Kuhn and A. Tucker, "Nonlinear Programming," in J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley 1951.

J. A. Mirrlees, "An Exploration in the Theory of Optimum Income Taxation," *Rev. Econ. Stud.*, Apr. 1971, *38*, forthcoming.

C. C. Morrison, "Marginal Cost Pricing and the Theory of Second Best," *Western Econ. J.*, June 1969, *7*, 145–52.

P. A. Samuelson, "Memorandum for U.S. Treasury, 1951," unpublished.

———, "The Pure Theory of Public Expenditure," *Rev. Econ. Statist.*, Nov. 1954, *36*, 387–89.

A Short-Term Econometric Model of Textile Industries

By Roger LeRoy Miller*

Short-term models have been devised for the purpose of answering both general and specific questions. These models have ranged from simple descriptions of an industry or an economy to specific estimates of certain parameters, such as output-factor elasticities or adjustment rates. The results presented here describe, in a somewhat unusual manner, two textile industries. The estimates of output-factor elasticities differ from those obtained by other investigators using single-equation estimation techniques. The model consists of a dichotomized (peak, off-peak) system of equations that attempts to capture relevant differences in decision making by firms during peak and off-peak periods in the production cycle. Investment in labor and in product inventories is considered, along with estimation of a production function and a demand-for-hours equation.

When an industry such as textiles is considered, production data clearly show definite seasonal fluctuations. In the analysis of decisions such as hiring and firing of workers, no adequate theory can ignore these (imperfectly) known fluctuations. During peak production periods most firms are faced with capacity constraints which do not exist during off-peak periods in the production cycle. During the latter the level of productive (efficient) employment of the hired work force may be considered as determined by "desired" capital-labor proportions of previous peak periods. That is, given the average full-capacity capital-labor ratio (presumed optimum)[1] and the rate of production during any off-peak period, the employment decision of the firm is not the total number of employees to hire, but rather the number to retain in excess of that dictated by the "optimal" capital-labor ratio and the rate of production. A prime decision is thus how much to invest in inventories of labor (reserve labor force) during seasonal partial-capacity production periods. Kenneth Arrow et al. (p. 18) recognized the need for an inventory approach to the demand for labor and Hollis Chenery considered similar "over-investment" behavior of firms as related to physical capacity. I extend this approach by focusing on inventories of labor and their interaction with other variables such as product inventories and production-worker hours.

Considering the relation between production (Q), sales (S), and finished-goods inventories (I),[2] a firm can hold production constant (never needing inventories of labor) while satisfying all changes in sales out of product inventories. Or it can vary production rates to match sales fluctuations. Or it can use a combination of both. If costs of inventory adjustment (C_I) are "high" relative to those of pro-

* University of Washington, Seattle. Many thanks are extended to Zvi Griliches for his numerous comments and criticisms and to an unnamed referee. Yoram Barzel, Tom Borcherding, Bruce Gardner, John Floyd, Leroy Hushak, Potluri Rao, Sherwin Rosen, Gerald Scully, Tom Simpson, and Arnold Zellner made helpful comments on an earlier draft of this paper.

[1] It is understood that this is an oversimplification; a firm producing steadily at peak output will employ fewer men and machines than it would have done at a seasonal peak.

[2] Changes in finished-goods inventories, $\Delta I_t = Q_t - S_t$.

duction rate changes, it may choose to vary production rates; then it must decide how to vary its stock of inputs relative to its production rate. Whenever the cost of of hiring, firing, and training exceed those of retention of inventories of labor (Z), the firm will hold the latter. Numerous authors have discussed the potential saving in turnover costs via the holding of a "reserve" labor force. (See, for example, Charles Holt et al. and Sumner Slichter.)

Walter Oi clearly demonstrated empirically the existence of labor as a "quasi-fixed" factor, as did Sherwin Rosen (1968); and Gary Becker theoretically treated this problem in human capital investment. Holt and M. H. David realized that the desired work force depends, *inter alia*, on finished goods' inventory and on forecasts of future sales, so that " . . . new vacancies need to be created in anticipation of future needs, ..." If these labor inventory vacancies were not filled, the firm " . . . would then face the following costly alternatives: working overtime, losing sales, or reducing [product] inventories to a level lower than desired" (p. 82). Holt (1969) implicitly suggested the existence of a reserve labor force when remarking that in response to a production increase " . . . labor productivity of the existing work force then increases by about 1 percent through *reducing slack time*, etc." (p. 137, emphasis added).

In what follows, I integrate the reserve labor theory into a system consisting of the following functions: production, peak-period demand for labor, inventories of labor, peak and off-peak demand for production worker hours, and peak and off-peak demand for finished-goods inventories.

I. The Production Function

A production relation should relate actual quantities produced to actual (as opposed to measured) quantitites utilized in the production process. The following Cobb-Douglas function is assumed:[3]

$$Q = A(E'H)^b K^c e^{pT} \tag{1}$$

where Q = quantity produced, E' = the number of technically effective employees, H = average hours worked per effective employee, K = total machine hours, T = a smoothly ascending time variable, and A, b, c, and p = parameters to be estimated.

Assume that E' is identically equal to measured labor E, whenever the firm is operating at full capacity. That is, during peak periods inventories of labor are nonexistent, E' becomes E in (1). The difference between E' and E is inventories of labor.

II. The Demand for Inventories of Labor

Given that firms desire a reserve labor force during slack production periods, what determines the desired stock? The firm typically has some idea of anticipated future changes in sales, especially if there are recurrent seasonals. For anticipated increases, there are two ways to assure that excessive backlogs do not pile up. One is to maintain larger levels of finished-goods inventories from which sales in excess of production can be met; the other is to maintain excess capacity in terms of men and machines so that the rate of production can be increased to match anticipated increased demand. It is thus obvious that, *ceteris paribus*, inventories of labor (and concomitant excess physical capacity) are substitutes for inventories of finished goods. That is, again *ceteris paribus*, desired reserve labor (Z^d) should be negatively related to finished goods in-

[3] For each short-run period, I assume negligible substitution between efficient units of labor and capital services, although not in the long run. The Cobb-Douglas function, with unitary substitution elasticity, is estimated with data from peak periods, thus allowing for long-run substitution (since most observations are four quarters apart).

ventories. This hypothesized substitution exists merely because these two inventories compete for the same purpose—to reduce the cost to the firm of backlogs.

It must be mentioned also that for any known sales course, a change in relative holding costs will induce a shift to a higher optimum level of the relatively cheaper inventory. For example, given the cost of holding I, any change in wages should result in a change in desired reserve labor. Naturally if there is little or no factor substitution, the wage rate, W, cannot affect the effective capital-labor mix. If, however, the costs of labor-stock adjustment such as recruitment, selection, training, etc., do not rise proportionally with the wage rate, then—given product-inventory holding costs—a rise in wages should induce a decrease in Z^d. Since I assume that the reserve labor force is not worked overtime, the relevant price here is the straight-time wage.[4]

It is true that one can think of situations where desired inventories will all move in the same direction. If a much larger permanent increase in sales is anticipated, the present value of the marginal revenue for holding reserve labor and for holding goods' inventories will exceed their marginal costs; hence both will be increased. Or, if the production cycle is perfectly in phase with the sales cycle but with a smaller amplitude, the corresponding inventory cycle will be positively related to the cycle of reserve labor. Since nothing in this model explicitly yields any indication of the timing of production relative to sales, it may be that the hypothesized negative relation between the two types of inventories will not show up. The problem is further complicated by the lack of explicit cost functions for the two inventories. Suffice it to say that the relative holding costs are changing and that if the data generate a negative relation between labor and finished-goods inventories (Z^d and I) and between labor inventories and wages (Z^d and W), then in some sense the idea of inventory substitution can be accepted.

[4] Note that though inventories of labor are measured in numbers of men, this may be an oversimplification since the reserve labor force may be composed of a fraction of every man on the payroll during off-peak production periods.

The demand function for Z^d should fall out of the relative cost function, which would include some measure of expected future sales. Without undertaking this task, I merely specify that Z^d is a multiplicative function of a few key variables for which data should be available:

$$Z^d = \Lambda(S^*)^{\delta} I^{r} W^{\phi} C_I^{\xi} \mathrm{T}^{\pi} \tag{2}$$

Desired labor reserve depends on a constant (Λ), anticipated future sales (S^*), product inventories (I), straight-time wages W, product-inventory adjustment costs (C_I), and a trend (T) added to take account of changes in secular demand.

Since I am assuming that firms do not hold inventories of labor during peak production periods, it is necessary to formulate an equation for peak-period labor demand.

III. Derived Demand for Labor During Peak Periods

Using an approach developed by Robert Ball and E. B. A. St. Cyr, assume a simple cost function:

$$C = W^*(EH) + F \tag{3}$$

where C = total direct costs net of materials, fuel, and variable capital user costs; W^* = the effective wage rate per man-hour; and F = fixed costs.

We note that W^*, being the effective pay per productive hour worked, will be "high" at lower hours per week worked and will gradually drop to a minimum when overtime rates are instituted, then will rise again. We can use a quadratic approxi-

mation to represent this relation between W^* and H:

$$W^* = g_0 - g_1 H + g_2 H^2 \tag{4}$$

For peak periods W^* will be a combination of straight and overtime pay and will thus be to the right of the minimum point mentioned above.

Putting (4) into (3) and adding the production function (1) constraint, the Lagrangian L is formed:

$$\begin{aligned} L = {} & g_0 EH - g_1 EH^2 + g_2 EH^3 + F \\ & - \lambda\left[Q - A(EH)^b K^c e^{pT}\right] \end{aligned} \tag{5}$$

Differentiating L with respect to E and H, setting the results equal to zero, and solving, gives static-equilibrium solutions for E and H:[5]

$$E^* = \left(\frac{2g_2}{g_1}\right) A^{-1/b} Q^{1/b} K^{-c/b} e^{-pT/b} \tag{6}$$

$$H^* = \left(\frac{g_1}{2g_2}\right) \tag{7}$$

Desired level of employment is a multiplicative function of a constant, output, capital, and time; whereas desired static level of hours is a constant.

I assume that due to adjustment costs there will be lags in the firm's peak-period response to changes in sales, hence in production and employment. Thus the standard partial adjustment mechanism is used with one *ad hoc* addition, a factor that takes into account how far actual hours vary from equilibrium hours (H^*). (E_{t-1} is last-period measured employment even if it was off-peak.)

$$\left(\frac{E_t}{E_{t-1}}\right) = \left(\frac{E^*}{E_{t-1}}\right)^q \left(\frac{H^*}{H}\right)^a \tag{8}$$

Substituting (8) into (6), and using (7), the peak-period labor demand function is:[6]

$$\begin{aligned} E = {} & \left(\frac{2g_2}{g_1}\right)^{q-a} A^{-q/b} Q^{q/b} K^{-cq/b} \\ & \cdot E_{t-1}^{(1-q)} H^{-a} e^{-qpT/b} \end{aligned} \tag{9}$$

Now actual peak labor demanded is a function of output, capital, lagged employment, production-worker hours, and time.

Desired hours (H^*) are shown to be a constant (7). This static solution needs modification for use in a short-run model.

IV. Demand for Production-Worker Hours

Hours worked provide the easiest and fastest variable for a firm to alter during transitory periods of change in production. Since little more than casual observation has been put forward on the demand for hours,[7] I apply in an *ad hoc* manner a modified partial adjustment mechanism to (7) to obtain the equation for peak-period hours demand:[8]

$$H = \left(\frac{g_1}{2g_2}\right)^j H_{t-1}^{(1-j)} Q_t^l Q_{t-1}^{-l} \tag{10}$$

Actual peak-period hours demanded are a multiplicative function of lagged hours and transitory changes in production Q_t/Q_{t-1}.

It is probable that a firm finding itself with large inventories of labor will wish to reduce paid-for hours faster for off-peak periods than otherwise, *ceteris paribus*. Thus we should find a negative relation between reserve labor, Z, and total hours

[5] Second-order conditions are easily verified. See Ball and St. Cyr and N. J. Ireland and D. J. Smyth for other derivations.

[6] Stability for (9) requires $0 \leq q \leq 1$ and $a > 0$.

[7] For an exception see Yoram Barzel and especially Rosen (1968, 1969).

[8] Stability requires that if $bl < 1$, then $j \leq (1 - bl)$ and if $bl > 1$, then $j \geq (1 - bl)$. As an unnamed referee has pointed out, since changes in H affect labor holding costs a formal link should exist between inventory and employment functions and the demand-for-hours equation. Equations (10) and (11) do take account of this.

paid-for per man per week. Therefore the reserve labor variable[9] is added to the equation (10) for peak-period hours demand to yield the off-peak period equation:

$$(11) \qquad H = \left(\frac{g_1}{2g_2}\right)^j H_{t-1}^{(1-j)} Q_t^l Q_{t-1}^{-l} Z^m$$

V. The Demand for Inventory Investment

The theoretical results obtained by M. C. Lovell are used here except for two modifications. All identities have been put into ratio form (which holds only approximately) in order to make the demand equation multiplicative. Also, the reserve-labor variable was added throughout in accordance with my hypothesis that Z^d and I are substitutes. For brevity, I will not go through Lovell's theory[10] but will present only the resultant inventory-investment equation:

$$(12) \qquad \frac{I_t}{I_{t-1}} = \mathrm{T}^h I_{t-1}^k (S_{t-1}^t)^n S_t^r (S_{t-1}^{t+1})^u \cdot (S_t^{t+1})^s (Z_{t-1}^t)^v Z_t^w$$

Variables subscripted and superscripted are anticipated ones: i.e., S_{t-1}^t is t'th period sales as predicted in period t−1.[11]

Equation (12) indicates that inventory investment I_t/I_{t-1} is a multiplicative function of lagged inventories I_{t-1}, t'th period sales as predicted using information up to period t−1 (S_{t-1}^t), actual t'th period sales (S_t), t+1 period sales as predicted in period t−1 (S_{t-1}^{t+1}), t+1 period sales as predicted in period t (S_t^{t+1}), predicted labor reserves (Z_{t-1}^t), and actual labor reserves (Z_t). The exponent terms h, K, n, r, u, s, v, and w, are to be estimated.

VI. Data and Estimation

Two textile industries are covered in this section: woolen weaving (221) and cotton weaving (223). The Bureau of Census publishes short-term data for these industries on capital services (machine hours), production, and inventories. Data on production workers and their hours are reported by the Bureau of Labor Statistics. Three variables had to be constructed: straight-time hourly wages (HW), expected sales (S^*), and inventories of labor (Z).

The wage rate is obtained by using data on average hourly earnings (AHE), average production worker weekly hours (AWH), and average overtime hours ($AOTH$).[12]

Expected sales are constructed with information on past sales by using a weighted moving-average technique developed by Holt et al. which utilizes seasonal and trend information. Since three weights needed specifying, it was decided to iterate on a three-parameter space until the constructed future sales series correlated most highly with actual sales, *ex post*. That is, over 200 predicted series were computed using different combinations of weights, and the "best" series determined the actual weights used.

A measure of Z was obtained by assuming that capital is utilized most efficiently when average hours per machine per week are at peak (full capacity). I assume also that the firm has some optimum ratio of men to machines that can be observed during these peak periods.[13] By plotting ma-

[9] As will be seen later, Z is an endogenous variable used as a proxy for Z^d in the simultaneous system.

[10] My full analysis will be furnished on request.

[11] Notice that (12) holds in its present form only for off-peak periods, since it equals zero when $Z=0$. To avoid this, assume that the reserve-labor variable is Z/Z^*, where Z^* is some long-run desired level, and that during peak period $Z/Z^*=1$. Then Z^* goes into the constant and the peak-period equation does not include the variable.

[12] Assuming that overtime pay equals 150 percent of straight pay, then:

$$HW = \frac{(AWH)\cdot(AHE)}{(1.5)\ (AOTH) + (AWH - AOTH)}$$

[13] Some firms have maintained that their most

chine hours on a time-series graph, capital-use peaks could readily be identified. (These are the peak periods for estimation of the full system.) The ratio of men to machines at these peaks was then computed and the result called "desired" ratios. A three-part moving average of the ratios was then calculated. For all off-peak periods the effective, or efficient, labor force, E_t, was obtained by:

$$(13)\quad E_t' = \frac{\text{observed machines in operation in period t}}{\text{average desired machine-men ratio}}$$

The reserve labor force was then obtained by:

$$(14)\quad Z_t = E_t(\text{observed}) - E_t'$$

This method allows for increases in labor productivity *pari passu* while still utilizing the assumption of negligible short-run capital-labor substitution. It is empirically accurate if the firm correctly anticipates peaks and has constant returns.

The final estimating form of the dichotomized system in logs with log-normal error terms added is:

Peak Period System

$$(1')\quad \log Q = \text{const.} + b \log (EH) - c \log K + pT + u_1$$

$$(9')\quad \log E = \text{const.} + \frac{q(1-c)}{b} \log Q + (1-q) \log E_{t-1} - a \log H - \frac{cp}{b} T + u_{9'}$$

$$(10)\quad \log H = \text{const.} + (1-j) \log H_{t-1} + l \log Q - l \log Q_{t-1} + u_{10}$$

$$(12')\quad \log \left(\frac{I_t}{I_{t-1}}\right) = \text{const.} - k \log I_{t-1} + r' \log Q + s \log S_t^{t+1} + u_{12'}$$

efficient rate of production is at less than 100 percent capacity.

Off-Peak Period System:

$$(2')\quad \log Z = \text{const.} + \delta \log S_t^{t-1} + \tau \log I + \phi \log W + \pi T + u_{2'}$$

$$(11)\quad \log H = \text{const.} + (1-j) \log H_{t-1} + l \log Q - l \log Q_{t-1} + m \log Z + u_{11}$$

$$(12'')\quad \log \left(\frac{I_t}{I_{t-1}}\right) = \text{const.} - k \log I_{t-1} + s \log S_t^{t+1} + w \log Z_t + r' \log Q_t + u_{12''}$$

The estimating equations differ from those presented earlier because of certain data limitations and problems:

(1) *Log K* was absorbed into *log Q* in (9) because of its almost perfect collinearity with the latter.

(2) High collinearity between the pairs S_{t-1}^t, S_t, and S_{t-1}^{t+1}, S_t^{t+1} forced me to delete the first of each pair from the inventory equation (12). The identity $\Delta I_t = Q_t - S_t$ (in ratio form) was used to eliminate S_t from (12). For simplicity it was assumed that $Z_{t-1}^t \cong Z_t$.

(3) Data were unavailable for constructing C_I in (2). It is assumed, though, that C_I varied little in relation to wages (W), so that a negative coefficient for *log W* in (2) should still be obtained.

(4) Notice also that (2) is in terms of Z^d whereas (2′) contains Z. I argued that Z^d is most importantly a function of expected sales, so that if computed expected sales are a good predictor of future actual sales, the use of Z will not introduce measurement error. Since my computed expected sales series were exceedingly well correlated with actual future sales, I consider Z an acceptable proxy for Z^d.

For the peak system, Q, E, H, and I_t/I_{t-1} are endogenous: the equations are overidentified. For the off-peak system, Z, I_t/I_{t-1}, and H are endogenous (Q being technically exogenous because of no capacity constraints). These equations are also overidentified.

TABLE 1—THREE-STAGE LEAST SQUARES RESULTS, QUARTERLY, PEAK PERIODS

Equation	Independent Variable	Woolens: Coefficient Estimate	Woolens: Coefficient ÷ St. Error	Cottons: Coefficient Estimate	Cottons: Coefficient ÷ St. Error
Production	1	−2.443	−0.932	−1.288	−0.606
	$\log (EH)$	0.655	2.233	0.427	3.466
	$\log K$	0.442	2.002	0.635	3.781
	T	0.004	2.645	0.001	1.231
Demand for Labor	1	1.391	3.333	−1.978	−0.743
	$\log Q$	0.451	7.989	0.465	5.491
	$\log H$	−0.018	−0.656	−0.010	−1.149
	$\log E_{t-1}$	0.418	3.236	0.477	0.836
	T	−0.002	−2.011	−0.001	−1.194
Demand for Hours	1	1.623	1.721	4.904	3.108
	$\log Q$	0.101	1.994	0.977	10.288
	$\log Q_{t-1}$	−0.099	−2.119	−1.043	−9.355
	$\log H_{t-1}$	0.744	6.762	0.650	3.510
Inventories	1	−3.776	−0.756	−1.281	−0.311
	$\log I_{t-1}$	−0.667	−2.123	−0.124	−1.792
	$\log Q$	0.811	2.468	0.233	1.249
	$\log S_{t-1}^{t+1}$	0.155	2.657	0.263	2.198

TABLE 2—THREE-STAGE LEAST SQUARES RESULTS, QUARTERY, OFF-PEAK PERIODS

Equation	Independent Variable	Woolens: Coefficient Estimate	Woolens: Coefficient ÷ St. Error	Cottons: Coefficient Estimate	Cottons: Coefficient ÷ St. Error
Reserve Labor	1	0.984	2.369	0.997	1.410
	$\log I$	−0.401	−8.488	−0.321	−6.817
	$\log W$	−0.146	−4.777	−0.100	−1.989
	$\log S_{t-1}^{t+1}$	0.374	2.001	0.349	3.687
	T	−0.0001	−1.114	−0.0002	−1.043
Demand for Hours	1	1.994	2.001	3.001	1.483
	$\log Q$	0.099	2.220	0.914	11.466
	$\log Q_{t-1}$	−0.116	−1.449	−0.989	−7.371
	$\log H_{t-1}$	0.668	2.798	0.569	1.982
	$\log Z$	−0.142	−1.398	−0.055	−1.117
Inventories	1	−4.987	−2.114	−0.993	−0.009
	$\log I_{t-1}$	−0.565	−6.428	−0.201	−2.444
	$\log Q$	0.771	5.976	0.276	2.119
	$\log S_t^{t+1}$	0.200	1.884	0.318	3.437
	$\log Z$	−0.605	−4.588	−0.115	−8.457

Tables 1 and 2 give estimation results using unrestricted three-stage least squares with quarterly data, 1960 to 1967.

VII. Analysis of Empirical Results

All regression coefficients had the theoretically correct signs, although several were not statistically "significant" at conventional levels. The coefficient of I in the demand for labor reserve was indeed negative, thus indicating that changes in relative costs during the sample period did elicit a substitution. The coefficient of the only included cost variable, W, was nega-

tive as expected. Both industries exhibited 9 percent increasing returns, but a null hypothesis of constant returns could not be rejected at conventional significance levels using a *t*-test.

Much recent empirical work (cited below) on implied output-labor elasticities was consistent with increasing returns to labor alone, a proposition that standard theory does not readily accommodate. My estimates of implied production function elasticities were consistent with decreasing returns to labor alone.

It is interesting to see whether the implied output-labor elasticities from the equation for peak labor demand conformed to those of the production function. That is, were the cross-equation restrictions satisfied? The direct-output labor elasticity b from the peak-period production function was 0.655 for woolens and 0.427 for cottons. Notice, though, that b appears as part of the coefficient of $\log Q$ in equation (9′). Since (9′) represents a difference equation, it is possible to obtain a long-run solution in terms of $\log Q$ and then to obtain the implied estimate of b.

The solution obtained is:

$$\log E = f\left[\left(\frac{1-c}{b}\right)\log Q, \ldots\right] \tag{15}$$

For woolens, inserting $c=0.442$ yields $b=0.765$. For cottons, inserting $c=0.635$ yields $b=0.785$. The first value is less than 10 percent off from the direct production

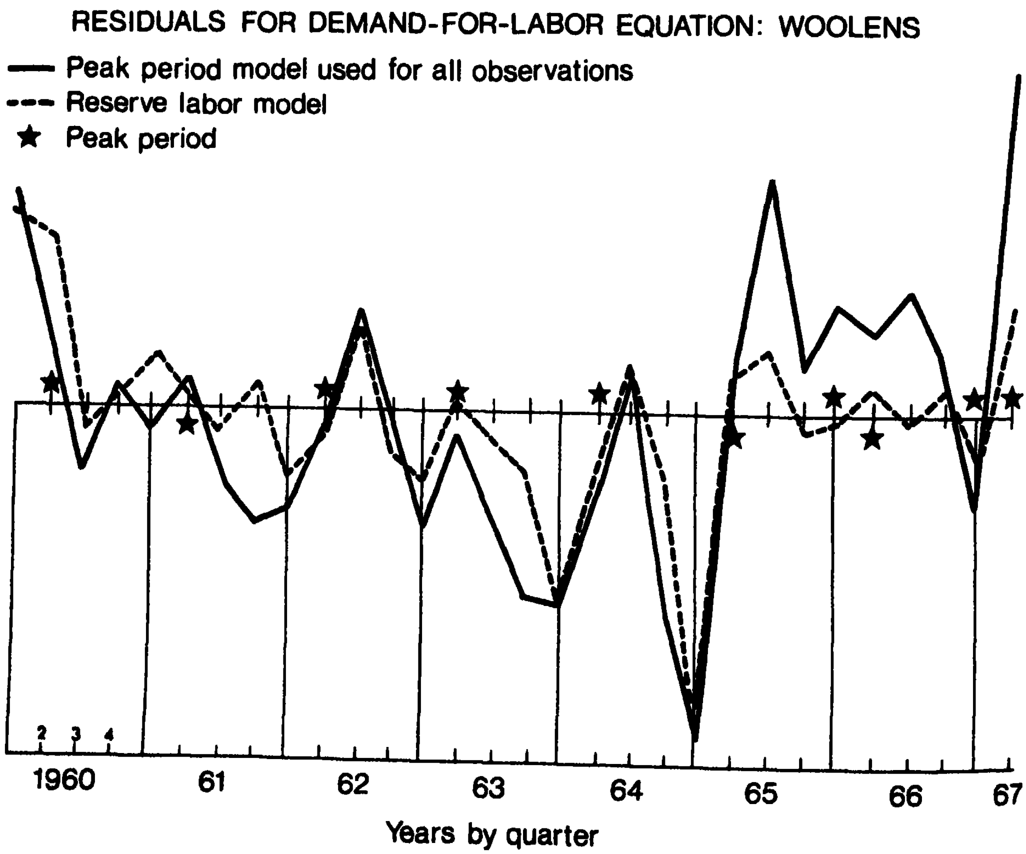


Figure 1

function estimate, but the second is substantially larger, although still less than unity. Such a small discrepancy for woolens indicates that a non-linear restriction was unnecessary during the estimation to obtain consistent cross-equation estimates of the output-labor elasticities. The same cannot be said for cottons. Alternatively, we can ignore the coefficient of *ln H* in (9′) and solve the resultant simple-difference equation for the implied coefficient of output: 0.77 for woolens; 0.89 for cottons.

We notice also that for the demand-for-hours equation, the magnitude of the coefficient of *log Q* is almost always approximately the same as that for *log* Q_{t-1}. Furthermore, the coefficients were always opposite in sign, thus indicating that the hours did in fact vary in response to transitory changes in output Q/Q_{t-1}, as postulated in Section IV. What is striking, though, is the large difference in these responses across industries; whereas a 1 percent change in Q/Q_{t-1} elicits about a 1 percent change in hours worked in the cotton-weaving industry, it elicits a mere one-tenth of 1 percent change in the woolen industry. This should be ample warning to researchers who attempt to use data on a more aggregated level than was done here.

Concerning the reserve labor variable, in the off-peak systems the coefficients of *log Z* were negative in the equations for demand for hours and demand for inventories, although not quite "significant" for the former. Moreover, in the reserve labor equation proper, the coefficient of *log I* was very highly significant and negative, thus lending more support to the proposi-

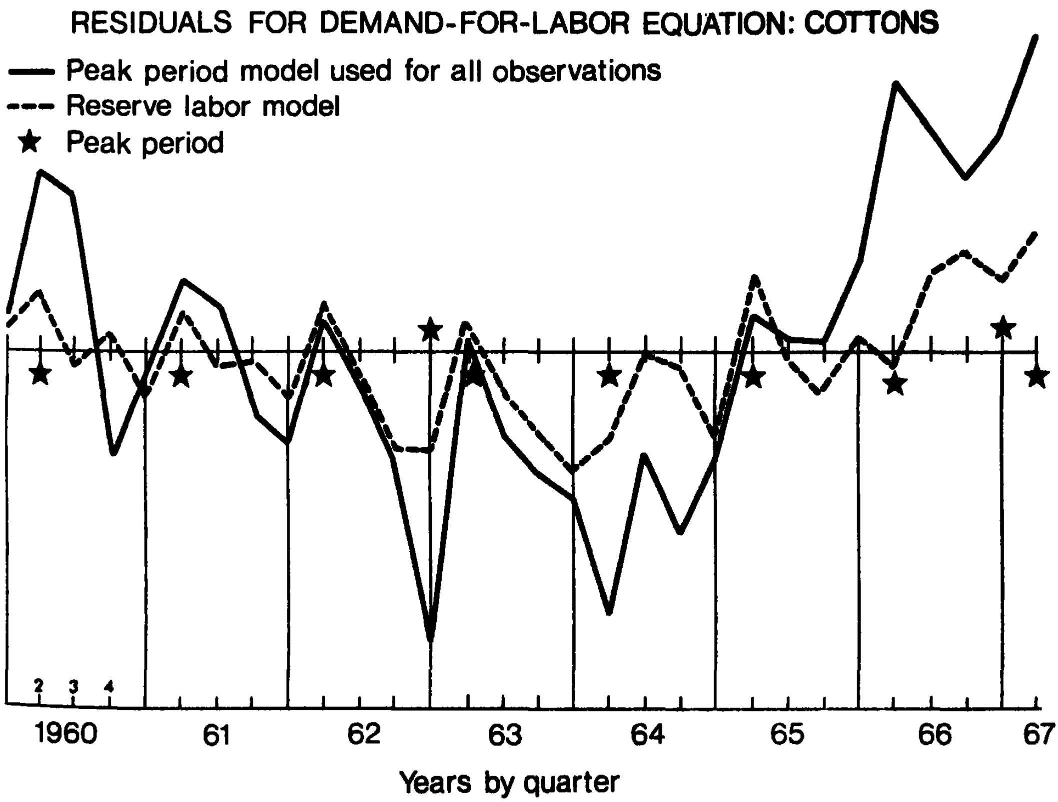


Figure 2

tion that inventories of finished goods are a substitute for inventories of labor.

Finally, I will attempt to assess the value of incorporating the idea of inventories of labor into the system. To do this, I reestimated the system using only the peak-period equations with all observations. The demand-for-labor equation (9′) (without H) has been used extensively in empirical work by Frank Brechling, Brechling and Peter O'Brien and others (Ball and St. Cyr, Smyth and Ireland, Robert Soligo). If the reserve labor theory is meaningful, the dichotomized system should be a better predictor of actual labor demand. I therefore computed predicted $\log E$ for each period from the estimates in the dichotomized system and from the estimates obtained from the peak-period system using all observations. Plots of the residuals are presented in Figures 1 and 2. Residual variances were computed along with Durbin-Watson statistics.[14] They are given in Table 3 below:

TABLE 3—RESIDUAL STATISTICS

	Residual Variance	D. W.
Woolens		
Peak-period system used for all observations	0.001973	1.0472
Reserve labor system with dichotomization	0.000952	1.662
Cottons		
Peak-period system used for all observations	0.002459	0.725
Reserve labor system with dichotomization	0.0005322	1.587

Use of the dichotomized system thus resulted in a reduction in residual variance of about one-half for woolens and about three-fourths for cottons. Moreover, much of the extreme positive serial correlation was eliminated by use of this system.[15] What still persists indicates that more improvement of the system is necessary.

REFERENCES

J. K. Arrow et al., *Studies in the Mathematical Theory of Inventory and Production*, Stanford 1958.

R. J. Ball and E. B. A. St. Cyr, "Short Term Employment Functions in British Manufacturing Industry," *Rev. Econ. Stud.*, July 1966, *33*, 179–207.

Y. Barzel, "The Determination of Daily Hours and Wages," mimeo., Univ. Washington 1969.

G. S. Becker, "Investment in Human Capital: A Theoretical Analysis," *J. Polit. Econ.*, Oct. 1962, Supp., *70*, 9–49.

F. Brechling, "The Relationship between Output and Employment in British Manufacturing Industries," *Rev. Econ. Stud.*, July 1965, *32*, 187–216.

——— and P. O'Brien, "Short-run Employment Functions in Manufacturing Industries: An International Comparison," *Rev. Econ. Statist.*, Aug. 1967, *49*, 277–87.

H. B. Chenery, "Overcapacity and the Acceleration Principle," *Econometrica*, Jan. 1952, *20*, 1–28.

C. C. Holt, "Improving the Labor Market Trade-off between Inflation and Unemployment," *Amer. Econ. Rev. Proc.*, May 1969, *59*, 135–46.

——— et al., *Planning Production, Inventories, and Workforce*, Englewood Cliffs 1960.

——— and M. H. David, "The Concept of Job Vacancies in a Dynamic Theory of the Labor Market," in *The Management and Interpretation of Job Vacancies*, Nat. Bur. Econ. Res. conference report, New York 1966.

[14] Comparison of residual variances is not strictly meaningful statistically because of the way predicted $\log E$ is computed in the dichotomized system. For computation of the D. W. statistic, the calculated residuals from the dichotomized systems were ordered seriatim from 1960 I to 1967 IV.

[15] Note that because of the inclusion of a lagged dependent variable in the demand-for-labor equation used for all observations, the Durbin-Watson statistic is biased toward two. This is not necessarily the case with the dichotomized system, where that equation contains a lagged dependent variable in only about one fourth of the observations.

N. J. Ireland and D. J. Smyth, "The Specification of Short-run Employment Models and Returns to Labor," mimeo. 1967.
M. C. Lovell, "Sales Anticipations, Planned Inventory Investment, and Realizations," in R. Fehher, ed., *Determinants of Investment Behavior*, Nat. Bur. Econ. Res. conference report, New York 1967.
W. Y. Oi, "Labor as a Quasi-fixed Factor of Production, *J. Polit. Econ.*, Dec. 1962, *70*, 538–55.
S. Rosen, "Short-run Employment Variations on Class-I Railroads in the U.S., 1947–63," *Econometrica*, July-Oct. 1968, *36*, 511–29.
———, "On the Interindustry Wage and Hours Structure, *J. Polit. Econ.*, Mar./Apr. 1969, *77*, 249–73.
S. H. Slichter, *The Turnover of Factory Labor*, New York 1919.
D. J. Smyth and N. J. Ireland, "Short-term Employment Functions in Australian Manufacturing," *Rev. Econ. Statist.*, Nov. 1967, *49*, 537–44.
R. Soligo, "The Short-run Relationship between Employment and Output," *Yale Econ. Essays*, spring 1966, *6*, 161–215.

Information and Frictional Unemployment

By Reuben Gronau*

An increasing number of economists discussing the trade off between unemployment and inflation have been trying recently to establish the microeconomic behavior giving rise to the Phillips curve.[1] A great many of these writers (see Armen Alchian, Charles Holt, Dale Mortenson, Edmund Phelps (1968), Melvin Reder) trace this behavior to the scarcity of information. According to this approach the unemployed worker in search of information on job opportunities lowers his wage demands as his search proceeds. Since the length of search is directly related to the level of unemployment, one would expect wage demands (and hence the change in the general wage rate) and unemployment to be inversely related.

This paper tries to reformulate George Stigler's theory of the economics of information (1961, 1962), to examine the factors affecting the job seeker's wage demands. It will be shown that even under conditions of perfect knowledge of future market conditions, maximization of expected income (or expected utility) will lead to a deterioration of the job seeker's real wage demands as the search continues. An increase in unemployment tends to quicken this process. Hence one can derive the Phillips curve without reliance on a lag in information (see Alchian), undercompensation in the setting of wage aspirations (see Holt), or money illusion. We examine the factors making the worker prefer voluntary unemployment to on-the-job search and the factors making him stop his search and drop out of the labor market altogether.[2]

The paper opens with a description of the optimum search policy. The analysis of the factors determining wage flexibility in Section II is followed by a discussion of the optimum combination of direct and indirect costs incurred through search (Section III) and its implication for voluntary unemployment. The concept of "despair" and movement out of the labor force are discussed in Section IV.

I. Optimum Search Policy

A job hunter embarking on search is entering a field of many uncertainties: he does not know what wage offers he will receive, when he will receive them, or how his wage will evolve through time. Even if one ignores the last factor, assuming, as I shall do for the rest of this paper, that on-the-job training and seniority result in an identical rate of increase of all subsequent wages, the job hunter is left with two unknowns. Both the timing and the

* Lecturer, the Hebrew University, Jerusalem. I am indebted to Gary S. Becker, Peter A. Diamond, Ruth Klinov-Malul, David Levhari, Jacob Mincer, and Tsvi Ophir and the referee for their useful comments, and to the Maurice Falk Institute for Economic Research for financial support.

[1] For a survey of recent literature, see Edmund Phelps (1969).

[2] While this paper was in the process of publication two other articles discussing the problem of job search and adopting a somewhat similar approach appeared. John McCall is interested mainly in the microeconomic aspect of the problem while Mortensen carries the analysis forward to derive a formal relationship interpreted as the Phillips curve. Both assume that, market conditions remaining unchanged, the job seeker's demands stay constant throughout the search, thus overlooking one of the more important features of the search process.

level of wage offers are subject to random distributions.

Given a job hunter with a finite time horizon, the time span can be broken into N discrete intervals, each small enough for the probability of more than one wage offer arriving during it to approach zero. The job hunter places a probability of p_n on the arrival of a wage offer during time interval n, the wage offer W_n being subject to a random cumulative distribution $F_n(W)$.[3] The variable W_n consists of money earnings (expressed in real terms) as well as of the psychic earnings the worker expects to derive from the job.

Search is a sequential process evolving through time. In each period the job hunter decides on his asking wage w_n. He stops his search if he receives an offer exceeding (or equal to) the asking wage $W_n \geq w_n$; otherwise he continues his search.[4]

The asking wage in period n is not independent of the asking wages the job seeker intends to charge in the future. Clearly, the length of search and the wage offer accepted depend on the asking wage and how it changes while the search continues. The crucial problem facing the job hunter is the formulation of an optimum strategy determining the asking wage at each point of time.

The asking wage will never fall short of any alternative option open to the job hunter. The optimum asking wage in period n should be set in such a way that any wage offer accepted during the period assures the worker of an income stream not inferior to the expected income stream he might have got had he continued his search. Specifically, let R_n be the present value of the income stream generated by a one dollar wage offer accepted in period n; then, given a job hunter who has no time preference, the utility derived from $R_n w_n$ should at least equal the expected utility yielded by continuing the search. Assuming that the job hunter is insensitive to risk, the necessary condition for the optimum asking wage requires the present value of the income stream generated by the asking wage, $R_n w_n$, to equal or exceed the income stream expected if the search is pursued beyond time n.[5] Let I_n denote the expected value of continuing the search beyond point n (i.e. the expected income stream), then the necessary condition for the optimum asking wage is $R_n w_n \geq I_n$ or $w_n \geq I_n / R_n$.

The alternative options open to a job hunter depend on his alternative earnings in the market and in the non-market sector and on how long each offer is open. The duration of offers determines, at least partly, the number of offers there are to choose from at each point of time. Since the job seeker will never accept any offer that falls short of the highest unexpired wage offer received in the past, he will tend to be more choosy, i.e., his asking wage will tend to be higher, the longer offers are open.[6]

The duration of offers determines the rate of obsolescence of the investment in information. It may vary from infinity down to zero. At the one extreme, the job seeker can return to any offer received in

[3] I assume that p_n is independent of $F_n(W)$, and that both are independent of the corresponding distributions in other periods. The assumption of a discrete time interval is not essential to the analysis and is adopted only for simplification. For an alternative approach, see Gustav Elfing.

[4] Note the difference between this model and Stigler's (1962). Stigler tacitly conceives of a worker who decides beforehand on the number of searches he will undertake, and who sticks to his decision regardless of the outcome of the search. For further analysis of the differences between Stigler's model and the sequential model, see P. Nelson's unpublished paper (some of which is included in his 1970 article).

[5] Both the assumptions of no time preference and insensitivity to risk can be relaxed at the expense of complicating the model.

[6] I assume that the fixed costs associated with search (both pecuniary and psychological) are sufficiently high to rule out the possibility of perpetual job changing, the worker moving from one job to a better paid one as it comes along.

the past whenever it suits him, with no danger of the offer having lapsed in the meantime.[7] At the other extreme, the job seeker is faced with an instantaneous decision—an offer not accepted is an offer lost.

This paper adopts the last assumption, assuming in effect that the search process has no memory. Every offer expires as soon as it is made; the worker has no alternative job offers to fall back on, and his asking wage in period n is, therefore, unaffected by past offers.

An employed job seeker will never accept any wage offer that falls below his current wages (where the wages include the money equivalent of the worker's psychic income). Similarly, an unemployed person will never consent to any wage offer that does not exceed the value of his marginal product outside the market sector (where this value includes the money equivalent of the utility derived from leisure) plus any unemployment insurance he may receive.

Let W^n denote the job seeker's alternative earnings in the market and non-market sector in period n ($W^n \geq 0$), then the asking wage w_n will never fall short of W^n. Since the worker has no incentive to offer his services at a price lower than the minimum value of the wage offers distribution $\underline{W}_n$ [where $\underline{W}_n = \text{Min } W$ satisfying $dF_n(W) > 0$], W^n will serve as the lower boundary of the asking wage, as long as it exceeds $\underline{W}_n$. When it does not, it will be replaced by $\underline{W}_n$.[8] Hence, the necessary and sufficient condition for an optimum asking wage is

$$(1) \qquad w_n = \text{Max}\,(I_n/R_n,\ W^n,\ \underline{W}_n)$$

Given the above assumptions, the optimum asking wage depends on W^n, $\underline{W}_n$, R_n, and I_n. The current wage W^n and the minimum wage offered $\underline{W}_n$ are parameters that can be assumed, as a first approximation, to be given exogenously.[9] In equation (2), the variable R_n is the present value of the income stream generated by a one dollar wage offer accepted in period n. Its value depends on the number of periods the worker intends to stay in his job (N−n), on the real rate of interest (r), and on the (real) rate of wage advances (β). In equation (2), the income stream is discounted to the initial period 0, and the worker is assumed to start working as soon as he accepts the offer.

The probability that a job offer will be received in period n is p_n. The probability that an offer received will be accepted is G_n where

$$(3) \qquad G_n = \text{Prob}\,(W_n \geq w_n) = 1 - F_n(w_n)$$

The probability that an offer will be received and accepted in period n therefore equals p_nG_n. Let E_n denote the average ac-

[7] This kind of assumption may have been at the back of Stigler's analysis when he discusses the returns of the search. Still, even in this case, the sequential model should be preferable to Stigler's static analysis.

[8] The notation used in the text implies that W^n does not necessarily belong to the distribution $F(W)$, i.e., the job seeker's current earnings may be outside the range of the job offers distribution (e.g., the job seeker is unemployed).

[9] As a first approximation, I ignore the interaction between the behavior of workers looking for jobs and that of employers looking for workers.

$$(2) \qquad R_n \begin{cases} = \dfrac{1}{(1+r)^N}\,\dfrac{(1+r)^{N-n+1} - (1+\beta)^{N-n+1}}{r-\beta} & \text{when } r \neq \beta \\[2ex] = \dfrac{1}{(1+r)^n}\,(N-n+1) & \text{when } r = \beta \end{cases}$$

ceptable wage in period n. The variable E_n is the expected value of W_n given that it exceeds w_n.

$$(4)\qquad \begin{aligned} E_n &= E(W_n/W_n \geq w_n) \\ &= \frac{1}{G_n}\int_{w_n}^{\infty} W dF_n(W) \end{aligned}$$

A job seeker entering period n places a probability of p_nG_n on finding within the period, a job that will yield him on the average an expected income of R_nE_n.

There is a probability of $1-p_nG_n$ that a job will not be found in period n. Search is a time-consuming process and thus may very well result in a decline of the job hunter's current earnings. Absenteeism or a decline in the time inputs in the home sector may reduce earnings to αW^n $(0 \leq \alpha \leq 1)$ as long as the search continues. Thus, if no job is found in period n, the job seeker will enjoy an income of $\alpha W^n/(1+r)^n = [R_n-(1+\beta)R_{n+1}]\ \alpha W^n$ during the period, and can look forward to an expected income of I_n in subsequent periods.

The variable I_{n-1} is the income stream the worker expects as he enters period n. Ignoring the direct costs of search, it equals

$$(5)\qquad \begin{aligned} I_{n-1} = p_nG_nR_nE_n &+ (1 - p_nG_n) \\ &\cdot \{[R_n - (1+\beta)R_{n+1}]\alpha W^n + I_n\} \\ & \qquad n = 1, \ldots, N, \end{aligned}$$

where $I_N = 0$.

Assuming the job seeker does not stop his search prematurely, one can insert the values of $I_n, I_{n+1}, \ldots, I_N$ in equation (5) to obtain

$$(6)\qquad \begin{aligned} I_{n-1} = R_n\alpha W^n &+ R_n(E_n - \alpha W^n)p_nG_n \\ &+ \sum_{k=n+1}^{N} R_k[E_k - (1+\beta)^{k-n}\alpha W^n] \\ &\cdot p_kG_k \prod_{i=n}^{k-1}(1 - p_iG_i) \\ & \qquad n = 1, \ldots, N, \end{aligned}$$

since $W^{n+1} = (1+\beta)W^n$ and $I_N = R_{N+1} = 0$. The job seeker entering period n expects with certainty to enjoy an income of $R_n\alpha W^n$ for the rest of his time horizon, and there exists a chance, measured by the probability

$$p_kG_k \prod_{i=n}^{k-1}(1 - p_iG_i),$$

that a new job will be obtained in period k, increasing his wealth, on the average, by $R_k[E_k-(1+\beta)^{k-n}\alpha W^n]$. The expected value of search is a function of the future search strategy. Note, however, that the weight of future decisions declines the further they are removed from the present. Decisions made in the distant future may, therefore, have only a negligible effect on the job seeker's behavior in the present.

II. The Factors Determining Wage Flexibility

How does the asking wage w_n change through time? Changes in the asking wage may be attributed to two sources: (a) changes that occur as part of the optimum search strategy, and (b) changes that result from a modification of the optimum strategy. A modification of strategy may take place because of unexpected changes in the labor market or because hitherto unknown information has become available in the course of the search.[10] For simplicity, it will be assumed that there exists perfect foresight on future conditions in the labor market [p_n, $F_n(W)$, etc.]. The worker knows the general shape of the wage offer distribution $F_n(W)$, but does not know which employer has a vacancy or the wage offer associated with each vacancy.

To obtain the optimum path of the asking wage, let us assume that I_n/R_n exceeds both $\underline{W}_n$ and W^n. By equation (5)

[10] This is essentially the case that Alchian has in mind. For a further discussion of this case, see Joseph A. Yahav.

$$\begin{aligned} w_{n-1} &= I_{n-1}/R_{n-1} \\ &= \frac{R_n}{R_{n-1}}\left[w_n + p_n G_n(E_n - w_n) \right. \\ &\quad \left. + (1 - p_n G_n) \frac{\alpha W^n}{(1+r)^n R_n} \right] \end{aligned} \tag{7}$$

$$n = 1, \ldots, N - 1$$

The asking wage in period n−1 is a function of current earnings, the asking wage in the next period, and the average acceptable wage in the next period.

The search must terminate within N periods ($I_N=0$). Hence, the job seeker cannot be choosy in the last period, his asking wage being equal to the minimum wage offer or to the alternative earnings, whichever is higher:

$$w_N = W^*_N, \tag{8}$$

where $W^*_n = \text{Max}\ (W^n, \underline{W}_n)$. Assuming $I_n/R_n > W^*_n$ for all n=1, . . . , N−1, the asking wage in period N−1 equals, by equation (5)

$$w_{N-1} = \frac{R_N}{R_{N-1}} [p_N G_N E_N + (1 - p_N G_N)\alpha W^N] \tag{9}$$

Combining this initial condition with equation (7), one may obtain a recursive solution for all values of w_n (n=1, . . . , N−2).

The exact nature of the asking wage path depends, of course, on the specific values of the parameters in equation (7). Still, one can trace some general features of the job seeker's behavior. The asking wage is determined by the ratio of the expected income from search (I_n) to the present value of the income stream generated by a one dollar wage offer accepted in period n (R_n). Both these terms decline as the search proceeds. However, assuming that the rate of job offer arrivals (p) and the wage offer distribution [$F(W)$] remain constant throughout the search, and that the job seeker enjoys no earning in the present ($W^n=0$), it can be shown that I_n declines at a faster rate than R_n, resulting in a fall of the asking wage over time.[11]

Other parameters remaining constant, the factor responsible for the decline in the asking wage is the finite time horizon. The job seeker's time horizon depends on the time he expects to stay on his next job. A job seeker embarking on search may expect (as I have assumed) to stay on his next job until a specified date (i.e., to a given period N), or, alternatively, for a specified length of time (i.e., for a given number of periods N−n). Given the first assumption, any prolongation of search will cut directly into the time spent on the next job. By the second assumption, the length of search does not affect the duration of employment on the next job, but, given the finite life span, will necessarily affect the time spent on some subsequent jobs. Either way, any prolongation of the search reduces the gains to be expected from the search.[12] The average period of employment in the United States is about 2.7 years, and the average period of unemployment is about one month (see Holt, p. 137), so that this reduction may be quite significant.

A major determinant of the asking wage path is the distribution of wage offers $F(W)$. It determines the asking wage threshold $w_N = \underline{W}$, and, given the asking wage in period n (w_n), the probability of acceptance G_n and the average acceptable wage E_n, two factors that are crucial in setting the asking wage in the preceding period (w_{n-1}).

Changing the scale of the distribution

[11] See Appendix.

[12] Mortensen explicitly assumes an infinite time horizon. McCall in essence adopts the assumption that the job seeker intends to stay on his job for a specified length of time but overlooks the effect of prolongation of search on subsequent jobs. Thus, both ignore the cost of elapsed time involved in search, and conclude that the asking wage remains constant throughout the search.

(i.e., multiplying the values of W and W^n by the same constant) does not affect the job seeker's behavior (i.e., w_n is multiplied by the same constant). However, given the center of location of the distribution (e.g., its mean), the expected gains from search increase with dispersion of the distribution (e.g., its standard deviation). Alternatively, given the dispersion of the distribution, the foregone earnings associated with continuing the search increase with the location of the distribution. Furthermore, the job seeker's optimum strategy may be influenced also by higher moments of the distribution, most notably, the symmetry of the distribution—the gains from search being greater the more skewed the distribution is to the left. Consequently, one would expect the deceleration of wage demands to increase with the positive skewness of the distribution and to decrease with its relative variation (i.e., its coefficient of variation).

The job seeker's chances of obtaining a new job depend, to a large extent, on the job offer stream. The job seeker's demands increase, therefore, with the rate of arrival of job offers (p).

The importance of the time of entry into the new job increases with the rate of interest (r) and the rate of wage advances (β). The higher the personal rate of discount, the lower the present value of any possible gains to be reaped by search, and, hence, the greater the urgency of getting a job fast. Similarly, the greater the remuneration for seniority and on-the-job training (i.e., the higher β), the greater the foregone earnings associated with continuing the search, and, hence, the greater the worker's tendency to accept lower wage offers at any point of time.

The incorporation of current earnings in the model cannot change the major conclusions reached above. Current earnings (W^n) increase the job hunter's boldness. The higher these earnings and the higher the percentage of earnings retained throughout the search (α), the smaller the loss of income associated with the search. Furthermore, when W^n exceeds $\underline{W}$ current earnings set the floor to any wage acceptable. Thus the higher the job hunter's current income, the higher his asking wage. An employed person can, of course, be more choosy than an unemployed one.

The search policy and market conditions determine the average length of search S_n.

$$S_n = p_n G_n + (1 - p_n G_n)(1 + S_{n+1}) = p_n G_n + \sum_{k=n}^{N} (k - n - 1) p_k G_k \prod_{i=n}^{k-1} (1 - p_i G_i) \tag{10}$$

The bolder the policy adopted by the job seeker, the longer he can expect the search to last. Given that the worker's demands fall as the search continues, so does the expected average length of search. Other things being equal, the length of search and the time horizon are inversely related.

The job seeker charges a higher wage the higher his current earnings, the higher the percentage of earnings retained throughout the search, and the greater the relative variation of the wage offer distribution. All these factors prolong the search. An increase in the rate of interest, on the other hand, moderates the job seeker's demands and curtails the duration of search. Finally, the higher the rate of unemployment, the lower the asking wage. It seems that the decline in wage demands does not offset completely the slower rate of job offers, however, and as a result a decline in p becomes a major factor in the prolongation of the search.[13]

How does an increase in the market rate of unemployment affect the wage demands of labor? An increase in the rate of unemployment, or a fall in the rate of vacancies, reduce the job offer stream and,

[13] For a more detailed analysis see Mortenson.

hence, accelerate the fall of the job seeker's demands. This tendency may be accentuated by several additional changes: (a) the increase in unemployment may be associated with a shorter employment period and, hence, a shorter time horizon; (b) the rate of interest facing the job seeker is lower the better he is equipped with assets (and, in particular, liquid assets). Prolongation of search may result in a deterioration of the unemployed worker's asset position, and in a subsequent increase of his personal discount rate; (c) when the job seeker is unemployed, W^n denotes his marginal productivity in the non-market sector and the money equivalent of his marginal utility of leisure. It has been suggested that a prolonged period of unemployment leads to a decrease of the marginal utility of leisure (see Hirschel Kasper) and, thus, to a decline of W^n; finally (d) the recession may be accompanied by a decline of the center of location of the wage offers distribution and by an increase in its positive skewness, "good" offers becoming more difficult to come by.

Thus, periods of increased unemployment are accompanied by a decline in the real wage demands of labor whether the recession is anticipated or not. The fall in the asking wage will be more abrupt when worsened employment conditions set in unexpectedly. However, one would also expect a gradual erosion of labor demands when the recession is expected, even if workers suffer from no money illusion, have perfect knowledge of the new market conditions, and there is no time lag between the change in prices and the change in wages.

III. Voluntary Unemployment

The discussion so far has given only a partial description of the factors bearing on the job hunter's optimum strategy. It depicts the factors affecting the length of search but does not explain the factors determining the intensity of search, i.e., the amount of resources devoted to search at each period.

The stream of job offers is a function of market conditions such as the market size and the state of employment, as well as of the amount of time and money resources applied to the search. Given the state of the market, the worker can accelerate or decelerate the rate of arrival of job offers by increasing or decreasing his direct or indirect costs. Direct costs of search include payments for employment agencies and advertisements, while the indirect costs consist mainly of foregone earnings.[14] If more time is spent on search, less time can be devoted to other activities. Since search seems to be more efficient when it takes place during regular working hours, an increase in time spent in search may very well come at the expense of time spent on the job, so that current earnings decline. The exact nature of the decline in current earnings depends on the amount of time diverted to the search, as well as on some specific factors, such as the employer's leniency towards absenteeism, the employee's opportunity to search during vacation, etc. An optimum search policy is expected to yield answers not only to the question how long to search, but also to questions about the optimum combination of inputs and the amount of resources that should be invested in the process.

Formally, the rate of arrival of job offers (p) is not a predetermined parameter, but an endogenous variable. It is a function of past and present investment of time and market inputs.[15] Let C denote direct costs,

[14] Payments to employment agencies may be in the form of a fixed share of earnings (say, one-half of the first month's salary). One can take account of this by an appropriate modification of R_n.

[15] I ignore any effect the investment in search may have on the shape of the wage offer distribution $F(W)$. An employment agency may serve as a screening device, passing only the best offers available. In this case, $F(W)$ becomes the distribution of the maximum values.

$$I_{n-1} = R_n W^n + p_n G_n R_n (E_n - W^n) + \sum_{k=n+1}^{N} R_k (E_k - W^k) p_k G_k \prod_{i=n}^{k-1} (1 - p_i G_i)$$

$$- \left[\frac{C_n}{(1+r)^n} + \sum_{k=n}^{N-1} \frac{1}{(1+r)^n} \left(\frac{C_{k+1}}{1+r} + (1-\alpha_k) W^k \right) \prod_{i=n}^{k} (1 - p_i G_i) \right. \tag{11}$$

$$\left. + \frac{(1-\alpha_N) W^N}{(1+r)^N} \prod_{i=n}^{N} (1 - p_i G_i) \right] \qquad n = 1, \ldots, N$$

and $(1-\alpha)$ be the percentage of current income lost, then[16]

$$p = g[(1-\alpha), C]$$

is the job offers production function describing the various combinations of time and money resources required to produce a given stream of job offers. The job hunter has to determine not only his asking wage at each period (w_n), but also the amount of time and market input ($1-\alpha_n$ and C_n, respectively) applied to the search. Put differently, he must choose the optimum set of policy variables $w=(w_1 \ldots w_N)$, $\alpha=(\alpha_1 \ldots \alpha_N)$, and $C=(C_0, C_1, \ldots C_N)$ so as to maximize I_{n-1} as shown in equation (11), subject to $p=g[(1-\alpha), C]$. The first part of equation (11) describes the expected gains from search and the second part (the terms in brackets) describes the costs.

The job seeker will combine the time and market inputs in such a way that their marginal rate of substitution in the production of job offers equals their relative shadow prices. If one assumes that the investment in search in period n does not affect the stream of job offers in subsequent periods (i.e., $\partial P_k/\partial C_n = \partial p_k/\partial(1-\alpha_n) = 0$ for any $k>n$), the optimum combination of direct and indirect costs is obtained when

$$g_{(1-\alpha)}/g_C = \frac{\partial I_n/\partial(1-\alpha)}{\partial I_n/\partial C} = (1 - p_n G_n) W^n \tag{12}$$

where $g_{(1-\alpha)}$, and g_C denote the marginal products of time and market input, respectively, in the production of job offers. Other things being equal, the tendency to substitute direct costs for indirect costs is stronger the higher the job seeker's current income (W^n), and the slighter the chances of accepting a new job (i.e., the smaller p and G). Thus, one would expect skilled workers to spend more on employment agencies and advertising than unskilled workers. Moreover, one would expect the job seeker to get more personally involved in the process of search (i.e., spend more time on search) as the process evolves and his asking wage declines (i.e., G_n increases).

Vacations being rare and employers' leniency towards absenteeism being limited, there may exist a maximum to the number of work hours the employee may miss, i.e., there may exist some lower bound to α (say $\underline{\alpha}>0$). If the optimum α_n falls short of $\underline{\alpha}$, the job seeker will prefer voluntary unemployment to on-the-job search (see Alchian).[17] Thus, given equation (12), one would expect voluntary

[16] The variables p, $(1-\alpha)$, and C can be interpreted as vectors. The rate of job offer arrivals in period n is a function of past and present investments in search $p_n = g(C_0, C_1, \ldots, C_n, \alpha_1, \ldots, \alpha_n)$.

[17] Both McCall and Mortensen ignore on-the-job search. However, this kind of search is far from negligible. Robert L. Stein found that in 1955, about 55 percent of all job changers did not suffer any unemployment.

unemployment to be more prevalent the lower the worker's earnings, and the more time has already been spent in search (the greater G_n). Finally, worker's tendency to quit their job in order to concentrate on search increases the greater the job offer stream p. Voluntary unemployment should therefore be more common in periods of rising aggregate demand, in particular if the rise in demand and profits is accompanied by a fall in employers' tolerance of absenteeism (i.e., a rise in α).

IV. Movement Out of The Labor Force

One of the assumptions made in Section II is that the job seeker does not stop his search prematurely. This assumption, ruling out despair, may prove too restrictive and, hence, will be relaxed in this section.

A job hunter stops his search altogether when he finds that the present value of his current earnings exceeds the expected value of the search. The present value of the earnings of a worker who stops his search in period n−1 is $R_n W^n$. Hence, the necessary and sufficient condition for despair is

$$(13)\quad R_n W^n > I_{n-1} \text{ or } W^n > I_{n-1}/R_n$$

When the rate of wage advances equals zero ($\beta=0$), $W^n = W^0$ for every n, and since $R_{n-1} > R_n$, a necessary condition for the termination of the search is

$$W^0 > I_{n-1}/R_{n-1},$$

that is, the job seeker does not stop his endeavors as long as his asking wage exceeds W^*_{n-1}.[18]

Let the rate of job offer arrivals (p) and the wage offer distribution stay constant through time (i.e., $W_n^* = W^*$ for every n). A job seeker hesitating whether to stop his search in period n−1 or whether to continue it for one additional period, will stop if

$$(14)\quad \begin{aligned} R_n W^0 > I_{n-1} = pG^* R_n E^* \\ +(1-pG^*)[(R_n - R_{n+1})\alpha W^0 + R_{n+1} W^0], \end{aligned}$$

where

$$G^* = 1 - F(W^*) \text{ and } E^* = E(W/W > W^*)$$

Ignoring direct costs, he will drop out of the search when the expected future gains of the search cannot compensate him for the income lost in the present.

$$(15)\quad \begin{aligned} &(1-\alpha)W^0 > pG^* \\ &\left[E^* - \alpha W^0) + \frac{R_{n+1}}{R_n - R_{n+1}} (E^* - W^0)\right], \end{aligned}$$

$\alpha < 1$ being a necessary condition for despair.

Put differently, despair sets in as soon as

$$(16)\quad \frac{R_{n+1}}{R_n} < 1 - \frac{pG^*}{1 - pG^*} \frac{E^* - W^0}{(1-\alpha)W^0}$$

where $p < 1$, $\alpha < 1$.[19]

The job seeker is more susceptible to despair the higher his current earnings W^0, the greater the loss of income due to search $(1-\alpha)$, the smaller the wage dispersion (i.e., the smaller $E^* - W^0$), the higher the

[18] The wage charged in period N equals $W^*_N = \text{Max}(W^N, W^N)$ and the asking wage in period N−1 should have been

$$I_{N-1}/R_{N-1} = \frac{R_N}{R_{N-1}} [p_n G_N E_N + (1 - p_N G_N)\alpha W^N]$$

However, given values of r and β that are sufficiently large (i.e., R_N/R_{N-1} being sufficiently small), and values of α, p_N, and G_N that are sufficiently small, there is no certainty that I_{N-1}/R_{N-1} exceeds W^*_{N-1}, and if it does not, the asking wage in period N−1 will equal the minimum acceptable wage W^*_{N-1}. Similarly, one can conceive of a case where the asking wage hits its floor in period n, and stays at the minimum level for the remaining (N−n) periods.

[19] The right-hand side of the inequality is independent of n. The validity of the inequality depends, therefore, on the left-hand side. R_{n+1}/R_n is a decreasing function of n. Thus, if the inequality holds for n, it will hold for every k>n. For example, when $r=0$ the worker drops out of the market as soon as

$$n > (N+1) - \frac{1 - pG^*}{pG^*} \cdot \frac{(1-\alpha)W^0}{E - W^0}$$

rate of interest (r), and the slower the rate of job offer arrivals (p).

When current alternative earnings fall short of the minimum wage offer (i.e., $W^0<\underline{W}$), the wage charged is the minimum wage offered; $W^*=\underline{W}$ and hence $G^*=1$ and $E^*=\mu$, where μ is the mean of the wage offer distribution. Rewriting equation (16), the search is stopped when

$$\text{(17)} \quad \frac{R_{n+1}}{R_n} < 1 - \frac{p}{1-p}\cdot\frac{\mu - W^0}{(1-\alpha)W^0},$$

$p<1$ being a necessary condition for despair.

An economic recession results in a dwindling stream of job offers (a decline in p), a downward shift in the job offers distribution (a decline in μ and E^*), and a possible increase in the personal rate of interest (r). All these factors have a discouraging effect on search. Employed persons will tend, therefore, to stick to their jobs, and the unemployed will be more easily tempted to drop out of the labor force altogether. The chances of finding a job and the difference between market earnings (μ) and the value of marginal product outside the market sector (W^0) are smaller for women than for men to start with. Women are the first to bear the brunt of unemployment, thus giving rise to the well-known "discouraged worker hypothesis."

On the other hand, when a husband loses his job, his wife may decide to enter the labor market. Having the husband at home reduces the value of the wife's marginal product, and the loss of this product suffered because of the search. The decline in W^0 and $(1-\alpha)$ increases the profitability of search, giving rise to the "added worker hypothesis."

V. Conclusion

In the preceding four sections an attempt was made to reformulate the theory of search and to apply it to some problems associated with frictional unemployment. It was shown that under constant market conditions, expected income maximization leads to a declining asking wage. Unemployment tends to accelerate the fall of job seekers' demands, and hence may lead to a Phillips curve-like relationship between wage change and unemployment. Costs of search minimization may lead the job seeker to substitute his own time for market inputs. At the extreme, the job seeker may prefer voluntary unemployment to on-the-job search to speed up the stream of job offers confronting him. A formalization of the concept of "despair" led to the analysis of the job seeker's incentives to join or quit the search (and labor force).

The search model can easily be adapted to describe employers' behavior in trying to fill vacancies. A more ambitious task is the application of the theory to the analysis of the expected period of employment, and hence, quit rates and discharge rates. Finally, one would like to carry the analysis from the individual level to the market level to analyze the determinants of the wage offer and the asking wage distributions as well as the actual distribution of earnings. These tasks still lie ahead.

Appendix

By equation (5)

$$\text{(1')} \quad I_{n-1}=p_nG_nR_nE_n+(1-p_nG_n)R_n \cdot\left[\left(1-(1+\beta)\frac{R_{n+1}}{R_n}\right)\alpha W^n+w_n\right]$$

Assuming that $W^n=0$ and that p_n and $F_n(W)$ remain constant over time

$$\text{(2')} \quad I_{n-1} = R_n[pG_nE_n + (1 - pG_n)w_n] \qquad n = 1, \ldots, N$$

It must be shown that $w_{n-1}>w_n$, i.e., $I_{n-1}/R_{n-1}>I_n/R_n$. Employing a proof by

induction: $I_N=0$, and $I_{N-1}=p_N G_N R_N E_N>0$; hence, the statement is true for $k=N$.

Let the statement be true for $k=n+1(w_n>w_{n+1})$, and show it is true for $k=n$. Let us assume that in period n, the job seeker charges an asking wage of $w_n'=w_{n+1}$ instead of the optimum wage w_n. The expected income in period $n-1$ is

$$
\begin{aligned}
I'_{n-1} &= R_n[pG'_n E'_n + (1-pG'_n)w_n] \\
&= R_n[pG_{n+1}E_{n+1} + (1-pG_{n+1})w_n] \qquad (3')
\end{aligned}
$$

By the induction assumption

$$
\begin{aligned}
I'_{n-1} &> R_n[pG_{n+1}E_{n+1}+(1-pG_{n+1})w_{n+1}] \\
&= R_n[I_{n+1}/R_{n+1}] \qquad (4')
\end{aligned}
$$

and hence

$$(5') \qquad I'_{n-1}/I_n > R_n/R_{n+1}$$

But I'_{n-1} originates in a sub-optimum strategy and hence $I_{n-1}>I'_{n-1}$. By equation (2)

$$
(6') \quad \frac{R_n}{R_{n+1}} = \begin{cases} \dfrac{(1+r)^{N-n+1}-(1+\beta)^{N-n+1}}{(1+r)^{N-n}-(1+\beta)^{N-n}} & \text{when } r \neq \beta \\[2ex] \dfrac{(1+r)^{n+1}(N-n+1)}{(1+r)^{n}(N-n)} & \text{when } r = \beta \end{cases}
$$

Hence $R_n/R_{n+1}>R_{n-1}/R_n$, and

$$
\begin{aligned}
(7') \qquad & I_{n-1}/I_n > R_{n-1}/R_n \Rightarrow \\
& I_{n-1}/R_{n-1} > I_n/R_n \qquad \text{Q.E.D.}
\end{aligned}
$$

The asking wage w_n is a decreasing function of n for any log-convex function R_n, regardless of the shape of $F(W)$.

By equation (2')

$$(8') \quad w_{n-1} = \frac{R_n}{R_{n-1}}[w_n + pG_n(E_n - w_n)]$$

Other things being equal, the asking wage increases with the rate of job offer arrival (p), the difference between the average acceptable wage and the asking wage (E_n-w_n) and R_n/R_{n-1}. The second of these terms is directly related to the wage offers distribution's relative variation and to its leftward skewness. The last term (R_n/R_{n-1}) is inversely related to both the rate of interest (r) and the rate of wage advances (β).

When $W^n \neq 0$ one has to correct (8') to obtain

$$
(9') \quad w_{n-1} = \frac{R_n}{R_{n-1}}\left[w_n + pG_n(E_n - w_n) + (1-pG_n)\frac{\alpha W^n}{(1-r)^n R_n}\right]
$$

The effect current earnings (W^n) have on the asking wage increases with α and decreases with p, w_n and R_n (i.e., with $(N-n)$). When R_n is sufficiently large, or α sufficiently small, the optimum strategy in the case of $W^n \neq 0$ approaches the policy adopted with $W^n=0$.

Finally, when p_n and $F_n(W)$ do not remain constant over time, one can always choose p_n small enough so that $w_{n-1}<w_n$. Similarly, one can assume that the distribution $F_n(W)$ lies to the left of $F_{n+1}(W)$ making the difference (E_n-w_n) sufficiently small to yield the same result.

REFERENCES

A. A. Alchian, "Information Costs, Pricing, and Resource Unemployment," *Western Econ. J.*, June 1969, *7*, 109–28.

G. Elfving, "A Persistency Problem Connected with a Point Process," *J. Appl. Probability*, Apr. 1967, *4*, 77–89.

C. H. Holt, "Improving the Labor Market Trade-off between Inflation and Unemployment," *Amer. Econ. Rev. Proc.*, May 1969, *59*, 135–46.

H. Kasper, "The Asking Price of Labor and the Duration of Unemployment," *Rev. Econ. Statist.* May 1967, *49*, 1965–72.

J. J. McCall, "Economics of Information and Job Search," *Quart. J. Econ.*, Feb. 1970, *84*, 113–26.

D. T. Mortensen, "Job Search, the Duration of Unemployment, and the Phillips Curve," *Amer. Econ. Rev.*, Dec. 1970, *60*, 847–62.

P. Nelson, "Information and Consumer Behavior," unpublished.

———, "Information and Consumer Behavior," *J. Polit. Econ.*, Mar./Apr. 1970, *78*, 311–29.

E. S. **Phelps**, "Money-Wage Dynamics and Labor-Market Equilibrium," *J. Polit. Econ.*, July-Aug. 1968, part II, *76*, 678–711.

———, "The New Microeconomics in Inflation and Employment Theory," *Amer. Econ. Rev. Proc.*, May 1969, *59*, 147–60.

M. W. **Reder**, "The Theory of Frictional Unemployment," *Economica*, Feb. 1969, *36*, 1–28.

R. L. **Stein**, "Unemployment and Job Mobility," *Mon. Lab. Rev.*, Apr. 1960, *83*, 350–58.

G. J. **Stigler**, "The Economics of Information," *J. Polit. Econ.*, June 1961, *69*, 312–25.

———, "Information in the Labor Market," *J. Polit. Econ.*, Oct. 1962, Supp., *70*, 94–105.

J. A. **Yahav**, "On Optimal Stopping," *Ann. Math. Statist.*, Feb. 1966, *37*, 30–35.

The Efficient Allocation of Subsidies to College Students

By STEPHEN A. HOENACK*

A current problem of public policy is to increase enrollments from disadvantaged groups in high quality colleges and universities. Even when students from disadvantaged groups are academically qualified, additional financial inducements are required to increase their enrollments, leaving less money to subsidize other groups. This paper attempts to illustrate the structure of tuition charges that will simultaneously achieve the objectives of total enrollment and the correct distribution of students from various income groups, given a fixed total subsidy available to the university.

The paper examines alternative ways of allocating subsidies among college students from different family income backgrounds. Since the University of California (the University) charges fees which are substantially below costs, all of its students receive an implicit subsidy.[1] The aggregate subsidy could be allocated in alternative ways by giving smaller subsidies to some groups of students and larger subsidies to other groups through differential tuition charges. Whether a particular alternative subsidy allocation increases or decreases the total level of enrollment depends on the nature of each group's enrollment demand curve.[2] The data for 1967 show that California high school seniors from different parental income groups were unequally represented at the University. Attendance was greater in the higher income groups. Estimated demand functions for attendance at the University show that a subsidy reallocation aimed toward a more equal representation from the income groups would increase total enrollment somewhat if the reallocation was small. However, beyond some point, a reallocation would bring about a smaller total enrollment, because successive movements towards equality become increasingly more costly in terms of reductions in the level of total enrollment.

We will hypothesize three alternative sets of objectives which can be achieved through a differential tuition policy, that the Regents of the University of California might have concerning the income group composition of the University's student body. The three sets of objectives have the common characteristic that the University has no preference for subsidizing students from one group more than students from any other group. They will differ in the University's willingness to make sacrifices in terms of the size of total enrollment for movements toward

* Associate professor, School of Public Affairs, University of Minnesota. The research underlying this paper was done at the University of California, Berkeley, and the paper was partly written at the Institute for Defense Analyses. The author gratefully acknowledges the support of both institutions. He wishes to thank George F. Break, Charles J. Hitch, John E. Keller, and Earl R. Rolph for encouragement and guidance during the study, and George A. Akerlof, Douglas C. Dacy, R. Daniel Gardner, Joseph A. Pechman, Henry M. Peskin, Rolf Piekarz, T. Paul Schultz, and an anonymous referee for invaluable suggestions on earlier drafts of this paper.

[1] Most of this subsidy is received from the state of California. This paper will not address the question of how large the subsidy should be.

[2] The existing empirical work on the private demand for higher education which is based on observed behavior of students has been done by Richard Ostheimer, Robert Campbell and Barry Siegel.

equal student representation from all groups. These objectives will be maximized subject to the constraints given by the size of the University's total subsidy and each group's enrollment demand function. The corresponding enrollments and fees for each group will be presented.

Section I describes the optimization procedure by which a university policy maker can use information about private demand for higher education in order to achieve his objectives with respect to total enrollment and the income distribution of students, given a fixed amount of subsidy. The use of this procedure is illustrated with the situation of the University funds in 1967. Section II summarizes briefly some of the results of a separate study (Hoenack) on private demand for attendance at the University. Section III shows the use of these results to determine the fees which would be charged and the enrollments which would result if the University were to use the optimization procedure outlined. I will present the results for the three alternative hypotheses about willingness to trade size of total enrollment for movements toward equal representation from income groups, defined by parental incomes of high school seniors.

I. Illustration and Application of the Optimization Model

This section presents a general description of a procedure for using information about private demand for higher education to help achieve objectives concerning the size and composition of total enrollment, given a fixed total subsidy for college students. Then it explains a specific application of this procedure to the situation of the University in 1967.

Illustration of the Optimization Model

Figure 1 illustrates the optimization model. Assume that there are only two categories of students, say students from families having incomes above and below the median level of family income in the population, and that their enrollments are measured A_1 and A_2, respectively. Assume that costs of educating students in the two groups are equal, and that demand for higher education is a positive function of income.

The area *Oab* represents the set of possible combinations of enrollments which can be achieved with a given subsidy. The curve, *ab*, is referred to as the enrollment possibility function (*EPF*). Since the higher income students have a higher attendance demand, the value *Ob* exceeds *Oa*; that is, the total enrollment would be larger if the entire subsidy were allocated to higher income students than to lower income students.[3] Both groups are equally represented at point *g*, where the 45° line intersects the *EPF*. At point *e*, where the slope of the *EPF* is minus one, total enrollment is at its highest possible level. In the portion of the *EPF* represented by *be*, there is no choice problem because reallocations of the subsidy increase both equality and the level of total enrollment. In the portion of the *EPF* represented by *eg*, any increase in equality reduces enrollment.

The curves labeled 1, 2, and 3 represent

[3] The *EPF* curve will in fact not touch the axes because demand is positive at a price which exceeds the university's cost of financing a college education. In this case the university will be making a net profit by taxing additional students in this category, thereby adding to the funds available for other students.

The convexity of the *EPF* curve can be demonstrated by calculating the second derivative of the function. The demand curve for A_i is $p_i = F_i\ (A_i)$ where p_i is tuition. The subsidy per student is the difference between c_i, the cost per student, and the tuition charged in category i. Thus, cost−subsidy$=F_i\ (A_i)$, and the *EPF* function is defined by the constraint $\sum(c_i - p_i)\ A_i = \text{constant}$. The slope of the *EPF* function is then

$$\frac{dA_1}{dA_2} = -\frac{[c_2 - MR_2]}{[c_1 - MR_1]}$$

where *MR* is marginal revenue. The convexity of the function is then easily shown.

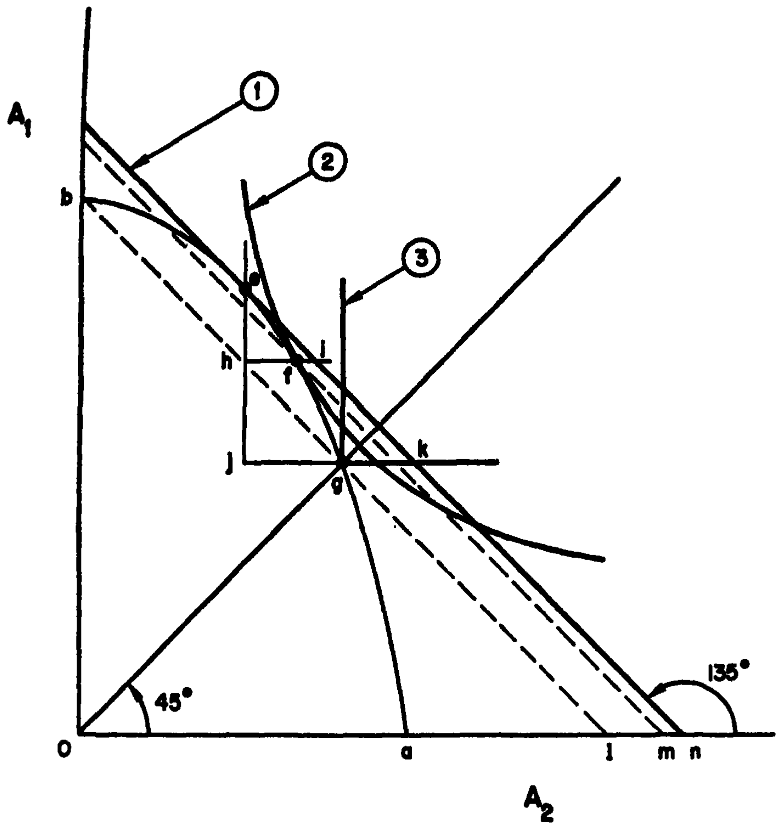


FIGURE 1

the indifference surfaces for three objective functions that might be used by policy makers to decide on the composition of the student body and total enrollment. With the first function the relative valuation of additional students in each income group does not vary with the relative level of enrollments from each group. While he has no preference for students in either group, the policy maker is concerned with enrollments. In the second function, the relative valuation of students from the two income groups depends on the relative enrollments from each group. When enrollments are unequal, the policy maker is willing to sacrifice more than one student from the group with higher enrollments for an additional student from the group with lower enrollments. This willingness to sacrifice size of total enrollment for movements towards equality increases with the degree of inequality. In the third function, only equal increases in enrollments from all groups are permissible. In this case, the policy maker's overriding concern is ensuring that the university's enrollments are drawn equally from each group, regardless of costs in terms of the size of total enrollment. Note that in the last two cases the policy maker is equally concerned if either group is disproportionately represented at the university.

The dashed lines in Figure 1 facilitate measurement of the sacrifice of total enrollment resulting from maximizing the second and third objective functions, instead of the first. The decrease in enrollments in either case equals the distance

measured on either axis, of the dashed line from the solid line (the indifference surface for the first function).[4]

Application of the Model

For the purposes of this study all 1967 California high school graduates were divided into four equal sized groups determined by their parental incomes. These quartile income brackets represent the broadest available socioeconomic measure of the population served by the University. The reason for taking equal numbers in the population brackets is to facilitate the formulation and interpretation of possible objective functions concerning the size and composition of enrollments. While the objective functions show value for total enrollment and varying degrees of value for the composition of total enrollment, they show no preference for any particular category of students. Only the position of the constraint provided by the enrollment possibility function can lead the policy maker to enroll a relatively large number of students from any particular one of the equally sized groups. A_1 exceeds A_2 in the case where the policy maker maximizes the first and second objective functions. The reason is the relatively lower cost from his point of view of enrolling students from the more well-to-do median group because of this group's relatively high private demand for higher education.

A further consideration is that the University's admissions standards are stringent. The illustrative calculations presented here are based on the assumption that admissions standards will not be changed as part of a policy to change the composition of the student body. Approximately the upper 12-1/2 percent of California's graduating high school seniors are eligible students within each population group. In Section II, I show that student representation at the University from these subpopulations depends on the costs of attendance. In addition, a high school student is more likely to prepare himself to meet the eligibility requirements to attend the University the lower are the costs of attending. Therefore reallocation of the subsidy affects the relative sizes of the eligible subpopulations.[5]

It would be possible to extend this analysis and consider admitting some students from low income groups who would require remedial education to become University eligible. However, the particular set of calculations presented here is concerned only with University eligible students.

The following discussion is a formal presentation of the optimization model used.

The Constraint (enrollment possibility function)

The enrollment possibility function expresses the set of feasible levels of enrollments from the four quartile parental income groups, given the size of the subsidy. Let

TS = total subsidy given to students from all parental income quartiles
p_i = tuition paid by students from the ith parental income quartile
A_i = number of students from the ith parental income quartile
c_i = the average cost of the university per student from the ith parental income quartile

Then[6]

[4] To see that this is so, consider measurements along the horizontal axis. In the case of maximizing the second function, A_1 is reduced by eh and A_2 is increased by hf. Therefore, total enrollments are reduced by fi which of course equals mn. Similarly, enrollments are reduced by ln if the third function is maximized instead of the first.

[5] The University ordinarily attempts to redefine its eligibility requirements in order to maintain its eligibility pool in the top 12½ percent of California high school seniors.

[6] An extension of this relationship can assist the policy maker who is interested in the choices between quality of education and the numbers of students (over)

$$TS = \sum_{i=1}^{4} (c_i - p_i)A_i \qquad i = 1, 2, 3, 4$$

The cost data in the relevant *EPF* for the University of California were based on the assumption that the c_i are the same for all parental income groups and that they do not vary within the relevant range.[7] The relationship between p_i and the A_i is the demand function of group i.

The Objective Functions

Let

α_i = the weight showing the relative importance of parental income quartile $i(\alpha_1=\alpha_2=\alpha_3=\alpha_4=\alpha$ where α is constant)

A_i = the number of students from parental income quartile i

B = level of achievement of the policy maker's objectives

Objective Function 1

$$B = \sum_{i=1}^{4} \alpha A_i = \alpha \sum_{i=1}^{4} A_i$$

In this case the policy maker is concerned with the total level of enrollments. The composition of enrollments is not relevant to his objectives.

educated. Assume that average cost per student (assumed to be the same for all types of students) is an effective index of quality. Also, assume that the college choices of students are influenced by quality. Then, $TS = \sum(c-p_i)A_i$ and $p_i = g^{(i)}(A_i, c)$ is the inverse of the demand functions. We then have $TS = \sum[c-g^{(i)}(A_i, c)]A_i$. One would expect that in some (low) range of values of c, increments of quality would bring about larger increments of p_i for given A_i. In this range the policy maker could simultaneously achieve larger enrollments and improved educational quality. Beyond some point, however, the p_i for given A_i would not increase enough to cover fully increases in c_i. In this range of c, the policy maker must choose between educational quality and size of total enrollments.

[7] Examples of possible differences in the c_i among groups of students are differences in instructional costs when students are grouped according to field of study and when students are grouped according to household characteristics, and remedial training programs for especially disadvantaged students.

Objective Function 2

$$B = \prod_{i=1}^{4} A_i^{\alpha} = \left(\prod_{i=1}^{4} A_i\right)^{\alpha}$$

In this case the policy maker is concerned both with the level and the composition of enrollments. His willingness to sacrifice size of the enrollments for movements toward equal representation from the four quartile groups is proportional to the existing inequality of representation from the groups.

Objective Function 3

$$B = [\min(A_1, A_2, A_3, A_4)]$$

In this case the policy maker attaches no value to any increase in enrollments unless there is an equal increase in enrollments from all groups. Operationally this objective function was maximized by maximizing the first objective function subject to the constraints that enrollment from all quartile parental income brackets had to be equal.

Maximization of the Objective Functions

The following Lagrangean expression was used to maximize each objective function subject to the enrollment possibility function:

$$(1) \qquad L = B(A_1, A_2, A_3, A_4) + \lambda[A_1(c - p_1) + A_2(c - p_2) + A_3(c - p_3) + A_4(c - p_4) - TS]$$

The first-order conditions for the constrained maximization are:

$$(2) \qquad \frac{\partial L}{\partial A_i} = \frac{\partial B}{\partial A_i} + \lambda\left[c - p_i - A_i\frac{\partial p_i}{\partial A_i}\right] = 0 \qquad (i = 1, 2, \ldots, 4)$$

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{4} A_i(c - p_i) - TS = 0$$

Equations (2) are a system of five equations with nine unknowns: the four A_i's

and p_i's and λ. Each p_i was expressed in terms of its corresponding A_i in the appropriate demand function, and thus eliminated.[8] The system of five equations was then solved for the A_i's and λ. The objective maximizing values of each A_i are then achieved by setting tuition equal to its corresponding value of p_i in the ith parental income group's demand function. Following a brief discussion in Section II of how the demand functions were estimated, the results will be presented in Section III.

II. The Estimated Demand for College Attendance

A major objective in estimating demand for attendance at the University of California is to estimate how costs affect enrollments for each of the parental income groups. A time-series model would have been inappropriate for this task because the available data are inadequate.

A cross-section model was used in which the observations were based on several hundred California high schools. Since tuition does not vary across such a sample of high schools, it was necessary to base differences in costs on transportation expenses. The monetary cost of transportation was estimated on the basis of the cash and time costs[9] of driving a car to and from college. The high schools were divided into two categories: those within and not within commuting distance of a University campus.[10] In the case of individuals who live within commuting distance of a campus, the transportation costs are the daily costs of traveling from home to college.[11] For individuals living beyond commuting distance of a campus, the transportation costs are the costs of traveling home on vacations. In regard to the demand behavior of eligible individuals living within commuting distance of a campus, only a sample of about sixty high schools within commuting distance of the University campus at Los Angeles was used. The estimated demand function based on this sample was assumed to represent the behavior of eligible individuals living within commuting distance of all university campuses. Separate demand functions for eligible individuals not residing within commuting distance of a campus were estimated for each campus except for the San Diego and Santa Cruz campuses. The demand for attendance at these two campuses was assumed to be the same as for the other University campuses combined.

For the observation on each high school the following data were used: number of graduating seniors attending each campus of the University, number of graduating seniors eligible to attend these institutions, incomes of families residing in the high school's attendance zone (through matching census tract data with maps of high school attendance zones), and wage rate

[8] The variable p_i in this discussion is the positive or negative level of tuition. The direct cost variable in the demand function is the total expense of going to college including tuition, if any. When this cost variable is solved in terms of A_i, it is necessary to subtract a constant representing all costs other than tuition from the solution when it is entered in place of p_i in the system of equations. In 1967 a fee was charged at the University of California to cover certain non-instructional services. The *EPF* for the University as defined above was the same whether or not this incidental fee is taken into account and the discussion ignores this fee for simplicity of exposition.

[9] The assumed time costs underlying the estimates in Table 1 were based on data from labor market surveys provided by the California Department of Employment. The results are not very sensitive to alternative assumptions about the valuation of travel time.

[10] Maximum feasible commuting distance is defined using assumed valuation of travel time, travel expense, and frequency and length of trips, as the distance at which it would be equally expensive to live at home and commute to college as to live on campus.

[11] In the demand study (see Hoenack) it was assumed that any individual living within commuting distance of a University campus who chose to undergo the extra expense of attending a campus away from home would not be affected in his choice to attend a University campus if costs were to change within a range which encompasses the variation of costs considered in the calculations in Section III.

and unemployment data for the part of the state in which the high school is located.[12] In addition, use was made of a computer program provided by the California Division of Highways which calculated the minimum driving distance and time between each high school and each University campus.

The following model was estimated. The jointly dependent variables were proportions of eligible high school seniors going on to attend each campus of the University of California. The independent variables were incomes of families in high school attendance zones, costs of attending each University campus and the nearest California State College and public junior college campuses, and wage and unemployment rates. (In the sample used for estimation of the attendance demand of eligible individuals living within commuting distance of U.C.L.A., there was no variation in costs of attending other University campuses or in wage and unemployment rates.) The remaining independent variables were interaction terms between each attendance cost and income. These interaction terms were used to estimate demand behavior separately for different income groups. For every estimated equation, the partial derivative of attendance proportion with respect to a change in the costs of attending all of the campuses was calculated for each income group. These partial derivatives were then weighted and summed to form the estimate of demand in each income group for attendance in the whole University system. Particular care was taken to make sure that the income variables used approximated the incomes of University eligible individuals. This was done by taking a sample of high schools for which both incomes in its surrounding census tracts and incomes of parents of University eligible students were known. These incomes were related to each other in a separate regression analysis. A detailed presentation of this demand study including the regression equations and the calculations made from them is found in my doctoral thesis.

The demand functions were estimated in two different ways in order to distinguish between short-run and long-run effects of changes in direct outlay costs. Lowering costs in the long run encourages students to become academically qualified and therefore raises the percentage of students eligible to attend the University of California, holding admissions standards constant. In the short run, however, a change in costs can only affect those students who have already qualified for entrance. The long-run demand for University attendance was obtained by adding to the short-run demand the estimated change in the number of eligible students resulting from a change in costs.

A logarithmic functional form was used in the estimation of the demand equations because it fit the data the best. Thus the estimated demand curves are steeper, the higher the cost of attending a University campus.

On the basis of the estimated relationships of the demand for the attendance at the University, the short-run elasticity of demand for attendance is −.85. The elasticity varies from −1.12 for the lowest quartile of parental incomes of all 1967 California graduating high school seniors to −.71 for the highest income quartile.

California state colleges are close substitutes for attending the University. Therefore, enrollment response would depend on whether there are simultaneous

[12] These costs were separately included in the regression equations only to help ensure that the equations were correctly specified. Foregone earnings should not be lumped with out-of-pocket costs because they represent in addition to opportunity costs of college attendance a potentially greater income while in college due to part-time and summer employment. Since college attendance is positively associated with income, the income effect of a rise in foregone earnings tends to offset its negative opportunity cost effect.

TABLE 1—PREDICTED EFFECTS OF INDICATED CHANGES IN ANNUAL COSTS OF ATTENDING THE UNIVERSITY OF CALIFORNIA ON 1967–68 UNIVERSITY ENROLLMENTS OF CALIFORNIA RESIDENT FRESHMEN

Income Bracket	Changes in Annual Costs	Percentage Change in Enrollment: No Changes in Costs of Attending Other Institutions		Percentage Change in Enrollment: Costs of Attending California State Colleges Change by 2/3 the Changes in Costs of Attending the University	
		No Effect of Costs on Eligibility	Costs of Attending College Affect Eligibility	No Effect of Costs on Eligibility	Costs of Attending College Affect Eligibility
$ 0,000–7,599	−400	30.04	36.78	17.87	24.61
	−100	6.87	8.24	4.08	5.45
	0	0	0	0	0
	+100	− 6.04	− 7.04	− 3.65	− 4.65
	+400	−23.03	−26.05	−13.78	−16.80
$14,500–& Over	−400	21.03	26.42	12.71	18.10
	−100	4.66	5.69	2.81	3.83
	0	0	0	0	0
	+100	− 3.81	− 4.47	− 2.57	− 3.23
	+400	−14.38	−16.06	− 8.90	−10.58
All Incomes	−400	24.31	30.24	14.51	20.44
	−100	5.46	6.62	3.26	4.43
	0	0	0	0	0
	+100	− 4.59	− 5.38	− 2.98	− 3.78
	+400	−17.45	−19.66	−10.67	−12.89

changes in costs of attending these other institutions. The elasticities were recomputed on the assumption that any percentage increase in direct outlay cost of attending the University is simultaneously matched by a two-thirds increase of that change in direct costs of attending the state colleges.[13] The recomputed elasticity of demand for freshman attendance at the University is −.51. Similar reductions in elasticities apply to all income quartiles on this assumption: The recomputed elasticity is −.68 for the lowest quartile and −.48 for the highest.

Table 1 presents the effects of $100 and $400 changes in direct outlay costs on proportions of eligible California high school seniors who attend the University of California for the lowest, highest, and all family income quartiles. The table shows short-run and long-run estimates.[14]

III. The Results

The system of five equations corresponding to each of the three objective functions was solved numerically. The results are presented in Table 2, under the alternative assumptions that costs of attending institutions other than the University of California are unchanged, and that the costs for any income group of California State Colleges change by two-thirds of any alter-

[13] The assumption of a two-thirds simultaneous change is used for illustrative purposes here; the figure was chosen because at one time it had policy relevance in California.

[14] The University is legally required to adjust periodically its admissions standards in such a way that the proportion of high school students who are eligible is constant. The long-run results presented here are based on the assumption that the aggregate number of eligibles is not held constant. Other long-run estimates were made on the basis of the assumption that eligibility standards are redefined so that the total number of eligibles is constant.

TABLE 2—ENROLLMENTS OF 1967–68 RESIDENT FRESHMEN AND CORRESPONDING TUITION CHARGES FOR ALTERNATIVE OBJECTIVES

A. Assumption: No Changes in Costs of Attending Other Institutions

Objective Function	Enrollment in Each Quartile Income Bracket								
	Quartile 1 Under $7,599		Quartile 2 $7,600–$10,499		Quartile 3 $10,500–$14,499		Quartile 4 $14,500+Over		All Income Groups
	A_1	P_1	A_2	P_2	A_3	P_3	A_4	P_4	A
Status Quo	2136	0	2577	0	4083	0	6157	0	14,953
Objective Function 1	2474	− 155	2836	− 84	4193	+ 21[a]	5863	+ 168	15,366
Objective Function 2	3035	− 488	3246	−307	4000	+ 97	4866	+ 474	15,147
Objective Function 3	3587	− 817	3587	−493	3587	+259	3587	+ 865	14,348

B. Assumption: Positive and Negative Tuition Charges in State Colleges Equal Two-thirds the Charges in the University of California

Objective Function	Enrollment in Each Quartile Income Bracket								
	Quartile 1 Under $7,599		Quartile 2 $7,600–$10,499		Quartile 3 $10,500–$14,499		Quartile 4 $14,500+Over		All Income Groups
	A_1	P_1	A_2	P_2	A_3	P_3	A_4	P_4	A
Status Quo	2136	0	2577	0	4083	0	6157	0	14,953
Objective Function 1	2454	− 230	2814	−136	4176	+ 14[b]	5876	+ 215	15,320
Objective Function 2	2973	− 699	3195	−449	3997	+120	4940	+ 639	15,105
Objective Function 3	3533	−1205	3533	−726	3533	+393	3533	+1060	14,132

[a] Defined by quartile income ranges of spring 1967 high school graduates. Negative tuition charges represent grants-in-aid.

[b] Positive charge corresponding to enrollments which are close to status quo enrollment levels are due to partial linearization of the demand functions.

ation in the tuition at the University of California. Thus, the second assumption implies a broader program in which the state reallocates its entire subsidy for all college students rather than only that part of its subsidy given through the University of California. Table 3 shows the effects of the alternative policies on enrollment composition.

If the University of California desires to maximize its total enrollment (Objective Function 1), it will change tuition to increase enrollments from the lowest two income quartiles by about 10 percent.

If the state's values are represented by the second objective function, it will allocate the subsidy to increase enrollments in the lowest two quartiles by 33 percent over the present levels. This allocation of the subsidy actually increases total enrollment by about 1 percent. The increase in the first and second quartiles is gained almost entirely at the expense of enrollment in the fourth quartile.

In the case of the third objective function, the state will reallocate its subsidy to increase enrollments in the lowest two quartiles by 50 percent, while decreasing enrollments in the highest two quartiles by approximately 30 percent. With the third objective function total enrollments are 4 percent lower than currently. In general, movements in the direction of equality are increasingly costly in terms of total enrollment.

The two alternative assumptions in

TABLE 3—PERCENTAGES OF ALL SPRING 1967 CALIFORNIA HIGH SCHOOL GRADUATES WHO ATTEND THE UNIVERSITY OF CALIFORNIA UNDER ALTERNATIVE OBJECTIVE FUNCTIONS BY PARENTAL INCOME GROUP

Objective Function	A: Percentage Representation from the Population of Persons in Each Quartile Income Bracket				
	$0,000–7,599	$ 7,600–10,499	$10,500–14,499	$14,500–& Over	All Incomes
Status Quo	3.4	4.1	6.5	9.8	6.0
Objective Function 1	3.9	4.5	6.7	9.3	6.1
Objective Function 2	4.8	5.2	6.4	7.7	6.0
Objective Function 3	5.7	5.7	5.7	5.7	5.7
	B: Percentage Representation of Enrollment of Persons in Each Quartile Income Bracket				
	$0,000–7,599	$ 7,600–10,499	$10,500–14,499	$14,500–& Over	All Incomes
Status Quo	14.3	17.2	27.3	41.2	100
Objective Function 1	16.0	18.4	27.5	38.1	100
Objective Function 2	19.9	21.6	26.6	31.9	100
Objective Function 3	25.0	25.0	25.0	25.0	100

Table 2 do not differ substantially in their corresponding enrollments, but they do differ in the changes in tuition charges necessary to maximize the objective functions. The reason for the large differences in tuition charges is that enrollment demand for the University of California is less responsive to changes in costs when costs of attending other suitable institutions change in the same direction.

Tables 2 and 3 reveal an interesting characteristic of the present tuition policy of the University of California. Evidently, the constrained maximization of the first objective function leads to enrollment levels in all of the income groups which are closest to the corresponding status quo enrollments. Therefore, the present policy in California comes closest to maximizing the first of the three stated objective functions. It appears that unlike the illustrative functions used in this study which show no preference toward any income quartile, the existing implicit objective function of the University does show some preference for the upper two income quartiles. This situation does not result from any conscious bias favoring relatively well-to-do students but rather because of the nature of the private demand for attendance. In order to achieve an explicit set of objectives the University should base its fees on the private demand for college attendance.

REFERENCES

R. Campbell and B. N. Siegel, "Demand for Higher Education in the United States," *Amer. Econ. Rev.*, June 1967, *57*, 482–94.

S. A. Hoenack, "Private Demand for Higher Education in California," unpublished doctoral dissertation, Univ. Calif., Berkeley 1967.

R. H. Ostheimer, *Student Charges and Financing Higher Education*, New York 1953.

Discrimination by Waiting Time in Merit Goods

By D. Nichols, E. Smolensky, and T. N. Tideman*

Perhaps the most ubiquitous of all urban problems is that the cities' public facilities—their roads, airports, shopping streets, license bureaus, schools, parks, beaches, pools, day care centers and public health clinics—are frequently crowded in ways that inflict time costs upon users. Waiting time does allocate public services, rationing them, as would money prices, according to the tastes, income and opportunity costs of consumers.[1] Time prices differ from money prices, however, since they appear relatively lower to persons with a lower money value of time. While such persons are likely to be considered more deserving, time prices have a defect: queues are a burden. It is alleged that some people, English housewives for example, enjoy a good wait. Despite such assertions, we will assume that queuing raises the cost of acquiring the good with which it is associated and that the burden from queuing is a deadweight loss. Time spent in a queue cannot be used productively.

The deadweight loss produced by a queue depends directly on the opportunity cost of time of those who wait. Thus when two individuals who value their time unequally wait in the same queue, they face different prices. This departure from the usual equilibrium conditions implies, in itself, that a queue of persons with different opportunity costs of time is inefficient. If trade were possible among persons who are waiting, or who might be paid to wait, this particular manifestation of inefficiency would disappear. Those with a low opportunity cost of time would resell to those with a high opportunity cost. Still, individual differences in the opportunity cost of time will affect the burden imposed by queues, because such differences determine who will wait and the length of the queue.

Money prices may be preferred to time prices because the revenues generated usually constitute an accurate signal. *Ceteris paribus*, it would be desirable to avoid the deadweight loss and to add to seller receipts. For these reasons, economists often recommend the imposition of user charges set equal to marginal social cost. However, a congestion charge in money is likely to be regressive in its effects, and several writers have agonized over whether the charge is justified simply because efficiency is increased in the process.[2] An alternative is to offer a public service at a wide range of money and time prices in a way that makes everyone better off. Our intention in this paper is to refocus the discussion of the efficacy of user

* Associate professor of economics, The University of Wisconsin, Madison; professor of economics, The University of Wisconsin, Madison; and assistant professor of economics, Harvard University. We wish to acknowledge the contributions of Samuel Morley, Jerome Rothenberg, Burton Weisbrod, and the participants in the 1968 Conference on Urban Public Expenditures where a preliminary version of this paper was presented. We are also grateful to that Conference and the National Institute of Mental Health for bearing a part of the costs of writing this paper. We were motivated to take up the problem of non-price rationing in the public sector after considering the provocative position taken by Julius Margolis (1968, p. 545).

[1] For a general discussion of time in the household production function see Gary Becker.

[2] See, for example, John Meyer et al. (pp. 334–41) and C. Sharp. Implicitly, these authors believe that there will not be any compensating redistribution.

charges on such an alternative program. At issue is not simply whether there ought to be a user charge in money at a single congested public facility, but rather how to achieve some constrained efficient allocation which is equitable.

Our single chain of argument will yield three major conclusions. First, we note that public services are frequently offered at a zero money price and then rationed by the waiting time required of recipients; furthermore, waiting time varies with the number of recipients. Since time is more equally distributed than money, this rationing device may be thought to be desirable because of equity considerations even though it is known to be economically inefficient. Since such equity considerations play no part in providing goods in the private sector, we conclude that public facilities are often congested for a reason in addition to those which lead to congestion in the private sector.

Our second principal conclusion is that queuing may be efficient. For public or private goods, queuing can be efficient if waiting by customers permits greater output. The efficient combination of queuing and money prices depends, of course, on the value of the customers' time. Queuing for public goods can also be efficient if there is a cost and a value to discriminating among recipients according to the opportunity costs of their time. The advantage of queues in this case stems from the fact that they enable us to charge different money prices to different groups without administrative cost. Facilities with higher money prices will have lower waiting times. A choice is thus offered the buyers which allows them to pay for the service with that combination of money and time which is cheapest for them. To low income buyers, combinations involving relatively higher time costs and lower money costs will be cheaper.

Finally, we conclude that the use of money prices to provide product differentiation may simultaneously improve equity and efficiency. We examine some equity issues that are inherent in alternative schemes for dealing with congestion and show that in some cases equity is improved in one dimension while it is worsened in others. We conclude by examining those characteristics of the social welfare function which must be known before one can unambiguously recommend the use of many money prices to partition the market for some otherwise homogeneous government service. While we restrict this analysis to differentiation by money price, we urge that the public sector consciously consider varying its conditions of sale along many dimensions. The chain of argument is also important because it leads to interesting questions, each worth considering in its own right quite separately from its relationship to the others. For example: Is the emphasis upon the need for marginal cost pricing in the public sector misplaced? Does the greater opportunity cost of time of the rich throw them into the private sector while leaving the poor in the public sector for selected consumption goods? Why do we provide merit goods and how do we determine the optimal capacity at a public facility providing that service? Does the reason for providing a merit good suggest the terms under which the good ought to be distributed? More generally, which rules for distributing merit goods are fair and which are not? We suggest a framework in which these questions arise naturally and we provide answers to some of them.

I. Congestion and Price Differentiation

Differentiating Products by Money Price

In the private sector, one way in which products are differentiated is by the time required to purchase them. One can spend time searching out merchandise at a dis-

count store, examining it, and waiting in line at the cashier. Alternatively, the identical commodity can be bought rapidly at a retail shop with the assistance of a clerk. The good will be more expensive at the shop, of course, since it costs the shopkeeper money to save the buyer's time. If competition prevails in the retailing industry, profits will be zero for both discount stores and retail shops. Customers with a high opportunity cost of time will prefer the shops while those with lower costs will use the discount stores. The equilibrium number of shops relative to discount stores will depend on the technical ability of stores to substitute time for money and on the distribution of buyers according to the costs of time. Assume that the only difference among the firms in an industry is the amount of time customers must spend purchasing their output. Competitive equilibrium in such a differentiated industry will have two requirements: First the profit rate in each productive process must be zero in the long run; second, each buyer must patronize that supplier which sells the commodity most cheaply, where the purchase price consists of the money price plus the value of time spent making the purchase.

To be in equilibrium the buyer must solve the following problem. Many ways exist to buy a commodity, some of which have high money costs but low time costs while others have high time costs and low money costs. A continuous frontier of such possibilities, *FF*, has been drawn in Figure 1. The buyer must choose that point from *FF* which minimizes his total cost. Following Becker, we assume the cost of the buyer's time to be his wage rate, and draw *AB* such that *AO*/*BO* is the buyer's wage.[3] The minimum cost point is *E*. Buyers with higher wage rates would prefer to pay with

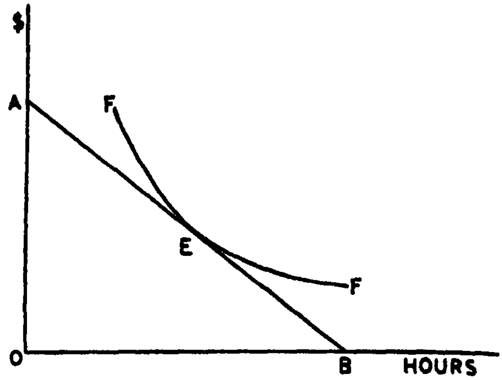


FIGURE 1

more money and less time. As the frontier is drawn, the technology of retailing is such that the store can save time for the buyer at some cost to itself. Buyers with high wage rates find it worth their while to save that time and pay the higher money price.

For producers, equilibrium requires that all points on *FF* yield zero profit.[4] Free entry guarantees this result in the long run. As long as average cost curves are U-shaped, an efficient competitive equilibrium results. This equilibrium is Pareto-efficient since the problem posed here is no different from the standard case of a firm in pure competition deciding what product to produce. Similarly, the individual's maximizing process is the usual one. We can view the purchase of each commodity as an activity with diminishing returns to labor. For an individual to maximize the return to his labor, his marginal minute in each activity must yield the same reward. If he faces a constant wage rate in one market, he must take part in all other activities until the marginal product of labor equals that wage rate in each activity. Thus the solution we have described is merely a special case of the general com-

[3] For *AEB* to be a straight line, we need to assume that the buyer faces a constant wage at which he may sell any amount of labor he chooses.

[4] If part of the frontier is non-convex, those points will not be observed as they represent inefficient processes.

petitive solution. Its efficiency depends on conditions which are well known.[6]

To illustrate the gains to buyers that result from differentiating a product by money price where previously differentiation had not existed, consider an example in which the money price is initially zero (to represent the conventional public provision of a service). The commodity is offered subsequently at both a zero and a positive price. The initial situation is represented by *A* on Figure 2. At the zero money price, its use is rationed by the *OA* man-hours in waiting time it costs to acquire it. Later the product is also offered at a second facility at money price *OB*. If no congestion resulted at the additional facility and therefore it took no time to buy the new commodity, it would be bought by all those whose wage rates exceed *OB*/*OA*. A more general result would involve some congestion, with point *C* ultimately describing the cost of the commodity at the new facility. Since we are assuming capacity unchanged at the old facility, the demand withdrawn from it would result in a new time price such as *E*. Those whose wages exceed *DC*/*DE* would find it cheaper to purchase at *C*; others would purchase at *E*. Thus from the buyers' viewpoint, differentiation by price which involves increasing money prices can lower the total cost of acquiring a service for *all* consumers, provided that capacity has been added.

It is also possible for providers to prefer differentiation even with increased capacity. With the added capacity, total costs to the government increase. The new buyers brought into the government facilities by option *C* and those who switch from option *A* to option *C* because it is a

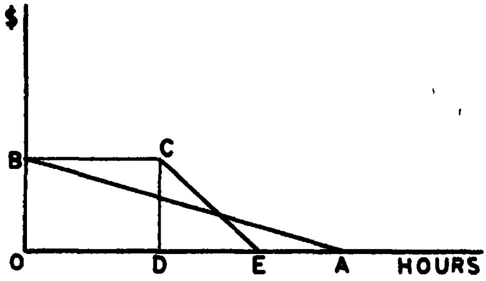


FIGURE 2

cheaper option, generate revenue for the government. It is an empirical question whether the revenue offsets the increased costs. It is at least possible. In the case where net costs increase, some social decision criterion must be consulted to see if the extra benefit is worth the extra cost.

Each additional point which might be added to Figure 2 would entail a set of calculations like that above. In the limit a continuum of money prices would be created. Varying queues would exist with the longest queue associated with the lowest money price.

Fluctuations in demand are an important source of the money-time trade off. Consider a group of privately owned facilities using the same technology to produce a product which is differentiated only by the money price charged. Assume that demand fluctuates through the day and that it is administratively impractical to vary money prices as demand varies, so that queues form from time to time. Where the money price is low, congestion is more frequent and more severe, so that the average amount of time necessary to make a purchase is higher. Where the money price is high, there will be less congestion on the average and a lower time price. Thus fluctuations in demand can produce a frontier like *FF* in Figure 1, simply because buyers respond to different money prices. If zero profits are still to exist at all points on the frontier, there must be some fixed factor which leads to

[6] Note that low wage buyers are not able to resell to high wage buyers because of the transaction costs involved. The ability to substitute cheap labor for expensive labor has already been exploited in the frontier, and is in fact, the reason for its very existence.

higher costs when the number of buyers is small. This would result, for example, from the existence of capacity that went unused at non-peak times. Buyers who wish to reduce the likelihood of queues must pay the costs of capacity which is not needed at non-peak times.[6]

Our concern is with publicly provided goods, but nothing said so far uniquely applies to them. Nor can public goods be introduced at this point by assuming insignificant long-run marginal costs, for if long-run marginal costs were insignificant there would be no congestion problem. To provide a rationale for public action it will be sufficient to assume that the commodity being provided is a merit good, i.e., that there is some public benefit from each unit sold. If such public benefits exist, and if those benefits do not depend on who consumes the commodity, then the efficient prices to charge individuals are given by a uniform downward shift of *FF* by the amount of the public benefit. For some publicly provided goods, however, merit value is related to characteristics of the consumer. Consider health services, for example. The poorer a person is, the more willing the public is to provide him with health services. It is this desire to differentiate among consumers according to income which undoubtedly provides the most satisfactory rationale for queues in the public sector, even though the time costs that result involve a dead-weight loss.

II. Queues that Deliberately Discriminate Among Merit Good Recipients

If the money cost of waiting time increases with the wage rate, any commodity that is rationed by a queue will be more expensive to those with high wage rates than to others. When confronted with alternative combinations of money and time prices, those with high wage rates choose the offering with a high money price and a low time price, while those with low wages choose the reverse. Thus if the public wishes to subsidize the money cost of a commodity to those with low wage rates only, they may offer it to all with a low (perhaps zero) money price, but offer such a small amount that a substantial queue results. To the low wage people, the money cost of the queue is minimal and they will receive a substantial benefit due to the lower money price. The high wage people will find the costs of the queue greater than the value of the money subsidy and they will not use the commodity even though its money price is low. Thus queues can be used to discriminate among users according to the opportunity costs of time. Even though the queue has an inefficient aspect in that the time of those who pass through it might have been used to raise total output without adversely affecting the buyer, nevertheless it is efficient overall if the alternative costs of discriminating—an equally effective means test—are higher. A queue is a decentralized way to discriminate according to the opportunity costs of time; it allows low wage people to select themselves as recipients of the money subsidy. Of course, if any alternative means of discrimination is cheaper, the queue remains inefficient.

Since queues may be the most efficient means of discrimination for some purposes, it is useful to discuss the nature of the problem faced by the government when determining the optimal length of a queue of a non-tradeable, non-storable commodity.[7] Queue lengths are determined in-

[6] This happens, for example, at supermarkets where product differentiation is effected by varying the number of cashiers employed. The more cashiers, the less often queues occur but the higher money prices must be.

[7] We have in mind here goods like visits to a health clinic, where the opportunity to reduce waiting time by buying more of the good each trip is virtually impossible.

directly, of course, since the control variables are the money prices charged and the quantities provided per time unit. The responses of individuals to these prices and to the resulting queues determine their lengths. For a given set of money prices, queues can be reduced by increasing the quantities of the services available.

The optimal quantity of the product to offer at any price is that amount at which the social cost of an additional unit just equals the social benefit. Comprehension of the relevant benefits and costs yields an understanding of the optimal solution.

Consider the problem of a government which wishes to subsidize the consumption of a commodity by low wage individuals and can offer a fixed subsidy to all potential consumers. For one consumer, the problem is simple. It is well known that such a subsidy must equal the value of the utility gained by other persons from the last unit of the commodity consumed by the individual. In Figure 3, *DD* represents an individual's demand curve for a commodity and it is known by the government. At each quantity, however, there are marginal external benefits to the general public which when added to the individual demand schedule produce the total marginal benefit curve *D′D′*. Given marginal production cost of *MC*, *AB* becomes the appropriate subsidy. With subsidy *AB*, the individual chooses to consume *OC* while he would choose *OE* in the absence of that subsidy.

If, as may be assumed, the same subsidy must be given to different individuals, there will be some welfare loss since the amount of public benefit, at the quantity he consumes, varies from one individual to another. For the individual pictured in Figure 4, the subsidy *AB* is not large enough to induce consumption of *OH*, the optimal amount, and a welfare loss represented by triangle *AGF* results. For some consumers, the subsidies will be too large

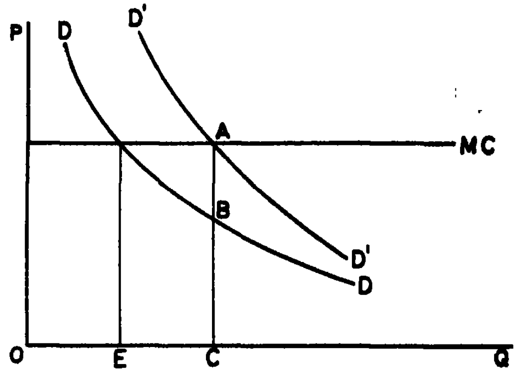


FIGURE 3

and corresponding triangles *AG′F′* will appear below the cost line, representing the fact that the value of the commodity to the individual plus its value to the government is less than its cost. If we continue to ignore the possibility of queues, the problem of the government is to select a subsidy scheme which minimizes the sum of triangles such as *AGF* and *AG′F′* over all individuals.

The use of queues to discriminate among users adds an additional source of welfare loss to that already represented by the triangles. This follows from the assumption that the opportunity cost of time is the wage rate, for by that assumption buyers would be indifferent between spending their time in the queue and paying out

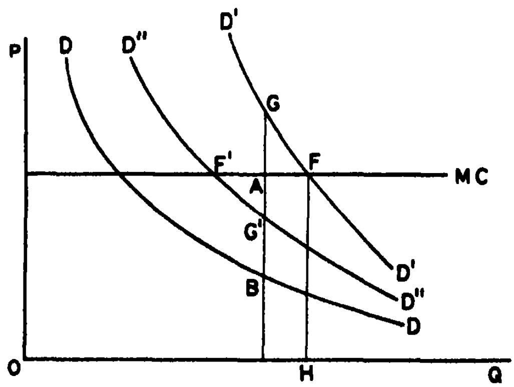


FIGURE 4

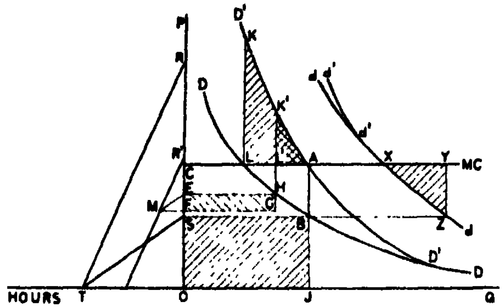


FIGURE 5

of the wages earned at their productive jobs. Nevertheless, the effect of the queues on the subsidies received by different individuals may reduce the total welfare loss. Queuing may lower the subsidy to those with high wages and raise it to low wage people. If this is the desired effect, it is possible that a welfare gain can result despite the additional deadweight loss implicit in the queue.

In Figure 5, the private demands of a low wage and a high wage individual are represented by *DD* and *dd*, respectively, while *D'D'* and *d'd'* represent private plus public demands for the same individuals. Note that the appropriate subsidies are *AB* and zero. If no subsidy is offered, the welfare loss is triangle *AKL*. If both are offered subsidy *AB*, the loss is triangle *XYZ*. Suppose that there is no uniform subsidy, and instead, the product is provided at a zero money price in the public sector and *MC* in the private sector, and a queue results in the public sector which requires an individual to spend *OT* man-hours to buy each unit. The individual's perception of the subsidized price will then be the money value of the waiting time. If the wage of the low wage person is *OS*/*OT* and that of the high wage person is *OR*/*OT*, then the money cost of the commodity in the public sector to the low wage person will be *OS*, while it will be *OR* to the high wage person. The time costs for the high wage person will exceed *MC*, the cost of the commodity in the private sector, so that the low wage person would buy from the public sector, while the high wage person would buy from the private sector. The only welfare loss results from the queue, a loss of *OJBS*, the value paid by consumers in waiting time that might, without additional cost to the buyer, have been used in production.

As drawn, *OJBS* is very large; an alternative pricing device exists which reduces this loss substantially. If the public sector charges a money price *OF* and offers a quantity that results in a queue which requires *FM* man-hours to acquire the commodity, the high wage person will still use the private sector although the price to him of the service in the public sector has fallen substantially, i.e., from *OR* to *OR'*, as long as the total cost to him, *OR'*, exceeds the marginal cost of the alternative, *OC*. The low wage person now pays *OF'* in money plus *FM* in time for each unit and consumes *EH* units. The total waste is *EFGH* plus *AK'L'*. A queue that minimizes the sum of this rectangle and triangle would be the most efficient scheme the government could offer when the costs of separating low from high wage people in any other way exceed that minimum sum. When the opportunity costs of individuals lie along a continuum, the government must select quantities of service to offer at various money prices in such a way as to minimize the sum of the rectangles and triangles over all individuals. That is, the sum of the wastes from standing in queues and from having different citizens consume too much or too little of a specific commodity must be minimized. In formal mathematical terms, this is a very messy problem; no significant solutions appear readily at hand.

A solution to a somewhat similar problem has been presented by William Vickrey. He discusses the provision of a nonmerit good (highway services) with time-

cost related to the number of users and shows that if there are constant returns to scale, competition will produce a continuum of money price-time price pairings in which the price that each person pays will cover the congestion costs imposed on others. Our problem is different and more difficult in that congestion serves as a sub-optimizing substitute for a means test rather than a way of making services available to more persons.

Differentiation of price may serve not only to vary the time cost to users, but also to vary other qualities of services. An obvious example would be the consumption of space at beaches, pools, and parks. Differentiating by price would leave some facilities less crowded than others, and allow thereby for differences in taste. One consequence might be to simply differentiate consumers by income class, which in at least some instances would be undesirable. Not all forms of product differentiation by money price are desirable, and we turn to a general consideration of their equity consequences in the next section. We conclude here by noting that examples of product differentiation can now be found in the public sector. Burton Weisbrod tells of a first-rate example of what is generally required. In San Juan there are "express" busses which run along the same routes as "local" busses and both are required to stop at the same places if their patrons demand it. Expresses, however, carry a higher price which tolls off customers and makes the express bus the more rapid travel mode. In this example the waiting time does not represent a deadweight loss since it is one of the necessary inputs for transportation, and is an example of Vickrey's model.

III. Equity and Money Price

The hard equity questions have been side-stepped until now since we have implicitly posited the existence of some social welfare function by specifying the public value attributed to each additional unit of consumption by an individual. Regardless of the form of that welfare function, certain systematic redistributions are implicit in any scheme involving the introduction of money prices into the public sector when they had not existed previously.

There will be a high but not perfect correlation between income and the opportunity costs of time. If we wish to treat those with equal income equally, we will find that the use of queues encourages too much consumption by those with low wages who have sources of non-wage income. And, of course, income may not separate those whom we wish to subsidize from those whom we do not wish to subsidize. Schemes which differentiate beneficiaries according to the opportunity costs of time are inappropriate if the society wishes to differentiate by other standards. The number and age of children, condition of health, or level of education may all affect the degree to which there is a public interest in enhancing the consumption of an individual, either in general or of a specific service. To the extent that these factors are present, queues will be an inefficient device for giving effect to such public interests.

Equity among income classes is also affected by the range of available prices. Suppose that a visit to a doctor is available at $0 plus 2-1/2 hours at the public clinic or $5 plus 1/2 hour in a doctor's office. The effective prices in money (the sum of money prices and time prices converted to money) for persons with different opportunity costs of time are shown by the solid line in Figure 6. Persons with time worth less than $2.50/hour use the public sector and pay 2-1/2 hours, while those with time worth more than $2.50/hour use the private sector and pay $5.00 plus 1/2 hour. Now suppose that an additional price is

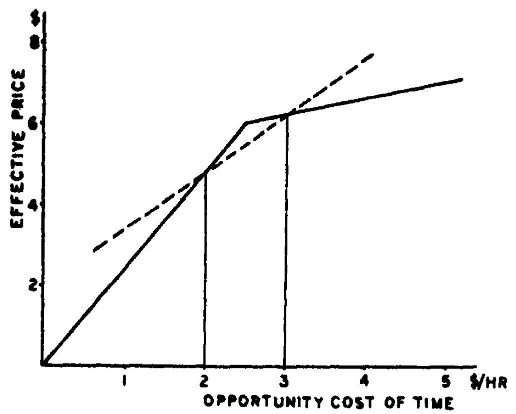


FIGURE 6

offered at the public clinic: $2.00 plus 1-1/2 hours. (This could be accomplished at a single facility by allowing each patient to specify the line he wished to join, and then calling patients in such an order as to preserve the relative lengths of the lines. The doctors would not need to know which line a patient came from.) The new effective prices, expressed in money, are shown by the dashed line in Figure 6. The beneficiaries are persons with opportunity costs of time in the range of $2/hour to $3/hour. The greatest benefits accrue to persons with time worth $2.50/hour.

The preceding discussion requires that the zero-priced facility continue to have the same time price. This is accomplished by adding an appropriate amount of capacity. If we were to merely institute money charges at some facilities where none existed previously, the queue lengths at the zero-priced facilities, serving the very poorest people, will increase. By varying capacity, any desired time price can be achieved.

It is difficult to set constraints on what constitutes equity. Nevertheless, it would be surprising if the corners of Figure 6 were consistent with maximization of any social criterion. Equity would seem to require a continuous variation in the nearly equal treatment of near equals, but a smooth schedule in Figure 6 would require a continuum of money-time price pairs, which might be prohibitively expensive. The practical optimum almost certainly involves undertaking some expense for the sake of greater continuity.

If we relax the assumption that the opportunity cost of time is measured by the marginal product of labor in the market, we can say little about equity. Relaxing the assumption does raise a pertinent question about equity, however, when the chosen policy is to expand the set of money price-waiting time pairings. Suppose, for example, that two individuals have the same opportunity cost of time in the labor market. For one of these individuals, the opportunity cost of time is indeed his marginal product of labor as valued in the market place. For the other, the opportunity cost of time is in performing motherly duties which she values at more than the market value of her time, so that she earns no wages. Introducing the set of money price-waiting time pairings benefits the worker but not the mother, who values her time so highly that she continues to pay the high money price.

Current practice which, by and large, offers public services on a first come, first served basis at a single money price (usually zero) poses its own problem in equity. As an appropriate test of an equitable rule we offer the following: *ex post* is the distribution of recipients of a service a random draw from the client population along each relevant dimension. That is, are there any systematic variations in characteristics between recipients and non-recipients which are arbitrary with respect to the purposes of the program? (See Morris Ginsberg, ch. 2.) Our expectation, which must still be empirically verified, is that current practice would not pass this test. Current practice imposes an arbitrary dis-

tinction: consumers will be differentiated from non-consumers by the opportunity costs of their time. The missing money price-waiting time pairings may even serve to discourage work effort from those at the margin between unemployment and employment at low wages. The value of waiting time may exceed the marginal product of labor at a full-time job. Casual employment may yield a higher total utility than the somewhat higher money income from full time employment for the poor hypochondriac—and hypochondria is a poverty-linked characteristic.

In summary, our policy proposal is to increase choice in the public sector through variation in money prices, which may be considered a form of third-degree price discrimination.[8] This proposal is not without equity problems. But without an explicit social welfare function on the one hand and a precise congestion function and production function for the service on the other, these conflicts are not resolvable.

Introducing Money Prices into the Public Sector: Some More General Issues

Failure to provide many alternative price-time pairings is not a widespread problem. Not only does it arise on a quite limited number of publicly provided goods, but it also probably affects the welfare of only a small portion of the income distribution. Many public facilities carry only one money price, usually zero, but the private sector supplies closely substitutable services at alternative price-time pairings. The two extremes of the income distribution are, therefore, probably getting the appropriate choices. The pairings offered, however, have a sharp discontinuity between the zero money price at which the good is fully subsidized by general taxes, and the minimum feasible money-time price pairing which just yields normal profits. Partial subsidies, with some money user charges, are lacking.

Partial subsidies have applications beyond those instances in which congestion is manifested by queues. Choice seems to be unduly restricted over the whole range of public service. One can go to the health clinic at zero money prices or the private doctor, the municipal golf course at zero money prices or the country club, the library at zero money prices or the second-hand book seller, the purely public or the purely private elementary school. The larger the public subsidy to any congested public service, the sharper will be the discontinuity in the price pairings offered. The general effect is to serve poorly those with a low but positive marginal product of labor.

Alternative price pairings for the same basic goods do exist throughout each metropolitan area. One source of these variations is residential segregation by income class. The charity health clinic that caters to the needs of domestics in high income areas is not likely to be congested. For other reasons, the public schools in the high income areas are not likely to be congested either. Political boundaries within the metropolitan area serve to permit individuals to collect according to income and their taste for public services, thereby producing a mix of money and time prices, but the commodity also varies.[9]

[8] Strictly speaking, third-degree price discrimination is an attribute of monopoly and an exercise of monopoly power. "A third degree would obtain if the monopolist were able to distinguish among his customers n different groups, separated from one another more or less by some practicable mark, and could charge a separate monopoly price to the members of each group" (A. C. Pigou, p. 279). We do not expect the government to exercise any monopoly power it may have simply to increase money receipts for its own sake.

[9] Charles Tiebout and Margolis (1957). However, as Paul Samuelson has pointed out, variance in income within neighborhoods reduces the effectiveness of product differentiation among municipalities in a metropolitan area (p. 377).

IV. Conclusions

The issues posed by this paper have not been completely resolved. We have done no more than highlight some relatively neglected facets of the problem of congestion at public facilities.

Because of the peak load problem, it is often efficient to have queues at both private and public facilities if the cost of varying prices exceeds the dead-weight loss of the queue. In addition, public services that are provided below cost often appear to be rationed through the use of queues. Such a rationing device is efficient only if the alternative forms of rationing are more costly than the dead-weight loss implicit in the existence of the queue. Queues are effective rationing devices because they impose a charge on the users in waiting time. Since the opportunity costs of time vary across people, the money cost of the queue will vary as well and for many people the costs of waiting will exceed the price at which the service can be bought in the private sector. Those with a high opportunity cost of time will find the money price in the private sector to be lower than the time price charged in the public sector. Thus the use of a queue rations a service exactly as if a money price were charged that varied directly with one's wage rate. Since the public sector often wishes to subsidize commodities in a manner that varies negatively with wage rates, queues can be an efficient device for singling out those it is desired to assist. This is true when alternative costs of discriminating exceed the dead-weight loss implicit in the queue.

In some cases, it may be desirable to charge many different money prices for the identical publicly subsidized commodity. Queues of different lengths will form with the shortest queues occurring at the facilities with the highest money prices. Individuals will then have a choice of paying for a commodity with various combinations of money and time, each choosing that combination which is cheapest for him. There may be substantial efficiency gains to be had from such differentiation.

Our proposal may produce serious equity problems that cannot be overcome.[10] Even if only the poor benefit, those with higher money income may benefit relatively more than those with lower incomes. If equity means the same treatment for all persons, it may not be possible to improve social welfare by increasing the number of money-time pairings. If, however, unequal treatment of unequals is equitable, which seems much more reasonable, then there are unexploited possibilities for improving social welfare. If offering the relatively better off among the poor services at both a smaller subsidy per capita than other poor and a smaller congestion cost is equitable, for example, then there should be a substantial increase in the set of the money-time price pairings offered the poor.

Taken together, the public and private sectors provide substitutable commodities at many alternative money-time price pairings. Those with a high opportunity cost of time will choose from the private sector, and the poor will choose from the public sector. Segmentation of the market will not extend to its technically feasible limits, however, unless governments offer income-in-kind at varying money prices. We think we have shown that there may be a high payoff in increased social welfare to ingeniously conceived expansions in the number of waiting time-money price pairings in the public sector. By extension, the payoff to increased welfare may also be extended by differentiating product along dimensions other than the money and time.

[10] The problems of extending choices on the supply side, thereby foregoing economies of scale, have not been discussed here.

REFERENCES

G. S. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

M. Ginsberg, *On Justice in Society*, Baltimore 1965.

M. B. Johnson, "On the Economics of Road Congestion," *Econometrica*, Jan.-Apr. 1964, *32*, 137–50.

J. Margolis, "Municipal Fiscal Structure in a Metropolitan Region," *J. Polit. Econ.*, June 1957, *65*, 225–36.

———, "The Demand for Urban Public Services," in H. S. Perloff and L. Wingo, Jr., eds., *Issues in Urban Economics*, Baltimore 1968, 527–65.

J. R. Meyer, J. F. Kain, and M. Wohl, *The Urban Transportation Problem*, Cambridge 1965.

A. C. Pigou, *The Economics of Welfare*, 4th ed., London 1960.

P. A. Samuelson, "Aspects of Public Expenditure Theories," reprinted in J. Stiglitz, ed., *The Collected Scientific Papers of Paul A. Samuelson*, Cambridge 1966.

C. Sharp, "Congestion and Welfare—An Examination of the Case for a Congestion Tax," *Econ. J.*, Dec. 1966, *76*, 806–17.

C. M. Tiebout, "The Pure Theory of Local Public Expenditure," *J. Polit. Econ.*, Oct. 1956, *64*, 416–24.

W. Vickrey, "Congestion Charges and Welfare," *J. Transp. Econ. Policy*, Jan. 1968, *2*, 107–18.

Optimal Mechanisms for Income Transfer

By RICHARD J. ZECKHAUSER*

The poor receive a sub-optimal amount of material goods through the economic processes of our society. Poor and non-poor alike seek to institute mechanisms that give something to the poor beyond what they get from the natural workings of the "competitive" system. That the non-poor share in this desire indicates that an externality is generated when the economic welfare of the poor is improved.

On a collective basis, the non-poor are willing to pay for the provision of this externality through transfer programs that benefit the poor. In the traditional terminology of externalities, a transfer program leads to an increase in the level of production of a good, in this case the economic welfare of the poor, which generates an externality.

This externality goes to all the non-poor. Thus, taking the non-poor to be the relevant group, this externality functions as a public good.[1] Once it is provided for one non-poor man, it is made available to benefit each non-poor man whether or not he contributed to its provision. Exclusion is not possible. As is usual with public goods, if individuals act in isolation a sub-optimal level of the good will be provided. That is why redistributive schemes of any magnitude are usually carried out by some collective body.[2]

In this paper, I am interested in the transfer program that would be chosen by a representative citizen who took into account that he would have to pay his proportional share of any collective effort. I assume that the citizen acts as a utility maximizer. His utility function is taken as given, so I need not be concerned with the moral or ethical values that underlie its structure. It is not of consequence here, whether the externality that enters his utility function is motivated by guilt, a sense of justice, or the desire to maintain a stable society.[3] What is important is the way a non-poor man benefits from improvements in the material well-being of the poor, the way the externality enters his utility function.

The elusive social welfare function is not dealt with here, nor are any other criteria relating to the welfare of the total society. All relevant welfare judgments are made on an individual basis by the participants in the society. It is probable, as I have suggested, that the providers of the scheme would choose to act collectively.

I conduct the analysis with the aid of a simple two-party model. One party is the

* Associate professor of political economy, Harvard University. I am indebted to a referee and the managing editor for many suggestions which have improved the paper. A preliminary and more detailed version of this paper was circulated in 1968.

[1] See James Buchanan for a recent discussion of the strategic elements of interactive situations in which groups provide themselves with public goods.

[2] Private charity cannot be relied upon to lead to a significant level of transfers to the poor unless there are substantial private benefits that return to the individual giver. Some individuals derive great inner satisfaction from giving, but for many others something more is needed. To this end, churches offer salvation to donors, universities name buildings for them, and the United Fund gives them stickers for their windows. The federal government encourages private charity through tax deductions.

[3] One need not be a Marxist to allege that poverty and societal stability do not go hand in hand. "Poverty is the parent of revolution and crime." Aristotle, *Politics*, Book II.

representative poor man. The other is the representative citizen; he is assumed to be non-poor.[4] The structure of the model is given by the assumed form of the representative individuals' utility functions and the permissible forms of income transfer schemes. After establishing this structure, I first examine the efficiency of negative income tax schemes. Then I turn to the major investigation, the problem of deriving the optimal income transfer scheme.

Quite obviously the optimal scheme in any particular circumstance will depend upon the parameters of the model. However, it is evident that certain predominant characteristics of optimal schemes can be discovered. To facilitate exposition, and indeed to make the optimization process computationally feasible, I sometimes find it helpful to employ numerical examples; in each case, the reader should understand that the specific functions and values of the parameters are selected to be illustrative, and that the major characteristics of the results can be generalized.

I. The Formulation of the Problem

The representative citizen asks himself the central question of this paper. *How should assistance programs to the poor be structured so as to maximize the utility function of the representative citizen?* I assume that considerations relating to assistance programs can be examined apart from the other decisions and allocations of the society, that the maximization of a subsidiary utility function relating to this single area will be consistent with the maximization of the representative citizen's comprehensive utility function.

My representative citizen, hereafter called A, has a utility function that includes three variables: The first variable, y, is the poor man's annual income after all transfers have been made. This variable, like all other money magnitudes in this paper, is measured in dollar units.[5] Assume for the present that A does not care about the poor man's expenditure pattern, that the poor man can best determine his own optimal consumption allocations, and that, other things being equal, the pattern that is best for the poor man is best for A. The second variable, h, is the number of hours the poor man works in a year. The third variable, c, is the annual dollar cost of the assistance program per poor man. The number of poor men in the society is known. Indirectly, then, c also gives the total annual cost of the program for the whole society. The total transfer to the poor, or perhaps A's share of it, may enter more directly into A's utility function. Given that these cost magnitudes are linear relations of one another, any one of them may be used as an argument of A's utility function. I use cost per poor man.

My representative poor man, hereafter called P, has a utility function that includes only the first two variables: his income and the number of hours he works. The utility functions for A and P are, respectively

$$(1) \quad {}_AU = A(y, h, c) \quad \text{and} \quad {}_PU = P(y, h)$$

Both functions are assumed to have positive partial derivatives with respect to y. The partial of A with respect to c is negative. It is not possible to prespecify the signs of the partials with respect to h for all values of h. Over the relevant range, however, we would expect the partial of P with respect to h to be negative. This means that P will choose to work fewer hours if he loses no income as a result. The partial of A with respect to h will be posi-

[4] In some societies the poor will exercise influence over the structure of assistance programs. In such societies the representative citizen is an amalgam of a poor and a non-poor man.

[5] To avoid the complications that accompany attempts to deal with incommensurate units, units of measurement are frequently omitted in this paper. Thus y here is given the value of 500, not \$500.

tive, at least until h takes on a very substantial value. Up through this point, other things being equal, A would like to see a reduction in P's leisure time.[6]

A's preference in this regard might reflect a paternalistic attitude. For P's sake, A might wish to provide him with the motivation, self-esteem, and beneficial experience that come from working at a job. Then again, A's preference may merely indicate his selfish desires. He may believe that additional work for the poor will reduce social unrest.

The major question of this paper, remember, is how should the representative citizen design assistance programs to the poor so as to maximize his own utility function. There is no general answer to this question. However, if we restrict the scope of our inquiry and specify the structure of the interactive situation, we can gain some insights. One possibility would be to treat this as a problem of trade in valued goods. A gives money to P in return for which P agrees to work some specified amount. The amount of the gift plus P's earnings would be the total of P's income. With full cooperation A and P can easily arrive at a Pareto optimal point. However, the type of cooperation needed to reach such an outcome might be impossible to achieve. Indeed, contractual arrangements of this nature might be considered unattractive.

Assistance programs are usually formulated on an impersonal basis. The transferring agent or agency, the representative citizen in this model, first draws up a general assistance program. The individual poor man then responds to the program in the way that is best for him. In terms of the model, A and P are engaged in a nonsimultaneous move, nonzero sum game. In the absence of cooperation, the situation has aspects of interdependence familiar to us from the theory of leadership oligopoly. There one firm, the first mover, markets a quantity assuming that the second firm will maximize its profits with respect to the market situation. This is the type of strategic planning that the representative citizen follows when he draws up the mechanism for income transfer that is best for him.

The model works in the following fashion. *A first sets up an assistance program; P then selects the number of hours he wishes to work in order to maximize his utility function subject to that program; A selects the program that gives him the highest utility payoff when P maximizes with respect to that program.* This formulation will be the basis for the analysis that follows.

II. The Negative Income Tax Plan

Economists have proposed the negative income tax (NIT) as a promising approach to the problem of redistribution. (See Christopher Green, Robert Theobald, James Tobin (1966, 1967), and James Vadakin.) Under the plan a poor man receives a stipend from the government, and this stipend is reduced as his income increases but by less than the full amount of the increase. The reduction in stipend represents a tax by the government on earned income. By contrast, income maintenance plans cut the stipend by one dollar for every dollar of earnings up to the base level. They form the limiting case of the NIT plan with the marginal tax rate equal to 100 percent.

It is unfortunate that these transfer plans have acquired the label negative income tax. They are negative only in the aggregate sense and apply a positive marginal tax rate to earnings. In this paper I follow convention to the extent that I label tax plans that are negative in the aggregate as NIT's. However, as I will

[6] Beyond some point, of course, A would like to see P have more, not less, leisure time. It is with this consideration in mind that we establish such measures as overtime laws and child labor laws.

discuss at length, it is most important to distinguish between plans that have negative marginal tax rates and ones, like those that have been widely discussed, that apply positive marginal tax rates to earnings. Hereafter, in this paper, unless otherwise specified, the term "tax rate" refers to the marginal tax rate on earnings.

The advantage of the *NIT* over income maintenance plans is that it retains some of the incentive to work. (See Peter Diamond.) At any income level some amount of marginal earnings will be retained. Figure 1 compares a positive marginal tax rate *NIT* plan with an income maintenance plan with the same base stipend. The recipient's indifference map is superimposed on the figure so it is possible to determine his maximizing response (the number of hours he chooses to work) for each of the two transfer plans.

In Figure 1:

wage rate = slope of *OA*,
base stipend = *OB*,
marginal tax rate under *NIT* = slope of *OA* minus the slope of *BD*,[7]
cost of the plan = the vertical distance from a relevant point on the two-segment line *BDA* to line *OA*.

Notice that northwesterly movements across the figure take the recipient to higher indifference curves.

As illustrated, the recipient would choose not to work under the income maintenance plan. Point *B* is on a higher indifference curve than any point on *CA*. With the negative income tax, he will choose point *E* where the two-segment line *BDA* reaches the highest indifference curve.

[7] Beyond point *D* regular tax rates may apply to reduce income after stipend and taxes.

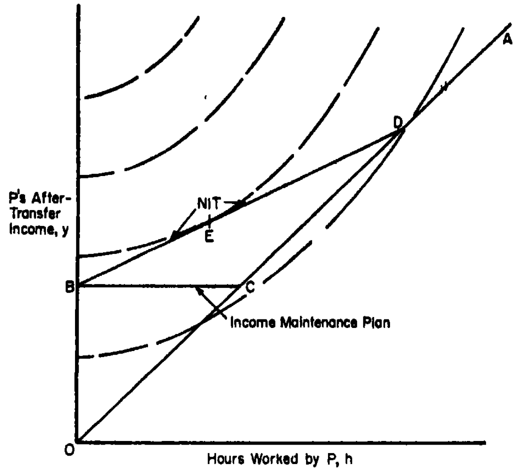


FIGURE 1. A NEGATIVE INCOME TAX PLAN COMPARED WITH AN INCOME MAINTENANCE PLAN

For the poor man, the *NIT* weakly dominates the income maintenance plan. For any amount of work it yields equal or greater income. Which plan is more costly to the government depends upon the amount that is worked under the two plans. The cost is given by the vertical distance from the income-hours point selected by the poor man to the wage-income line, *OA*. The more effective the *NIT* work incentive, the lower is its relative cost.[8]

[8] The poor man might work more under the income maintenance plan than he would with the *NIT*, even though the partial derivative of his utility function with respect to work is everywhere negative. This will be the case if the best point for the poor man on *AC*:1) is on a higher indifference curve than point *B* (the assumed best point on *BC*) and therefore will be selected with the income maintenance plan; 2) is on a lower indifference curve than the best point on *BD* and therefore will not be selected with the *NIT*; and 3) lies to the right of the best point on *BD* and thus represents a greater work effort. Unless leisure is an inferior good, 3) will hold whenever 1) and 2) are satisfied.

The case in which these three conditions are satisfied is quite different from the case in which leisure is an inferior good, so that raising its price (decreasing the marginal tax rate on income) decreases the amount of leisure consumed (increases the work effort). If it were an inferior good, then as the price line pivoted upwards from *BC* to *BD*, higher utility levels would be accompanied by greater work effort. That is, the *NIT* would produce a greater work effort than the income main-

An Example with the Negative Income Tax

The possibilities A considers are all negative income tax plans with constant marginal tax rates. A selects the base stipend, s, and the tax rate, t. The transfer plan thus established determines P's after-transfer income, y, as a function of the hours he works, h, and his wage rate, w. This income equals the base stipend, plus P's wage income multiplied by one minus the tax rate. That is,

$$y = s + hw(1 - t) \tag{2}$$

Given the transfer plan, P picks h, the number of hours he wishes to work, in order to maximize his utility function, $P(y, h)$. In the terminology of oligopoly theory, P reacts to the s and t selected by A.[9] This response by P determines the cost of the plan to A. That cost, c, is the difference between P's after-transfer income and his wage income if he works h hours. In algebraic terms we have

$$c = [s + hw(1 - t)] - hw = s - hwt \tag{3}$$

A manipulates s and t to maximize his utility function, $A(y, h, c)$. Here h is determined, given that P will make his maximizing response, and y and c are defined by (2) and (3) as functions of this maximizing h.

To investigate this interactive situation, we must employ some hypothetical utility functions. The selected functions have the partials specified in the discussion of (1), the expression that gives the general form of the utility functions. Two of P's indifference curves derived from his hypothetical utility function are presented in Figure 2.

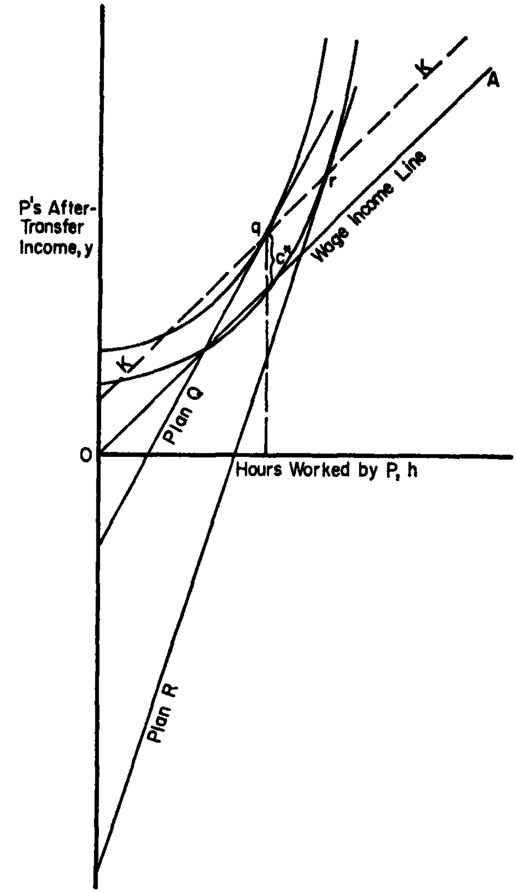


FIGURE 2. TRANSFER PLANS OF EQUAL COST

Each of the possible income transfer plans A could establish can be represented as a straight line across the diagram, which is called a plan line. The base stipend, s, for the plan will give its vertical intercept, the amount P receives if he works not at all. The slope of the plan line indicates the amount of extra income P receives for an additional hour of work. Thus, this slope minus the wage rate gives the marginal tax rate, t, for the plan.

---

tenance plan. This contradicts the hypothesized observation.

If the recipient works more under the income maintenance plan than he would under the NIT, it must be because he is functioning on the CD segment of the income maintenance plan. On this segment the price of leisure is greater than it is with the NIT; the inferior good explanation would not be relevant.

[9] P's reaction function has two arguments; it cannot be represented in a two-dimensional diagram. However, every s, t pair determines a unique plan line. As Figure 2 illustrates, P's reaction to a plan line can be shown in a two-dimensional diagram.

Assume that A established the transfer plan represented as plan Q in the figure. P will choose point q, the point along the plan line that offers P the highest utility, the point where it is tangent to one of his indifference curves.

The cost of this plan to A will be the difference between P's after-transfer income and his wage income. It is shown as c^* on the figure. The line KK shows all combinations of after-transfer income and hours worked for P which cost A the amount c^*. Any plan that leads P to make a maximizing response that lies on KK will cost the same amount. Observe that P's maximizing response to plan R is at point r, a point on KK. This means that plan R will cost A the same amount as plan Q.

Characteristics of the Optimal Plan

To determine the characteristics of A's optimal plan, let him first deal with the single family of plans that cost c^*. Given a choice between transfer plans Q and R, A would decide whether he preferred point q or point r. The two points cost the same. Point r yields P a higher after-transfer income; A values P's income positively. Point r also represents a greater work effort for P; A would likely regard this as an asset as well. With two areas of advantage and none of disadvantage, A would choose plan R over plan Q.

What properties accompany this preference? The base stipend for plan R is significantly more negative than the one for plan Q. The marginal tax rates for both plans are negative. This is illustrated by the fact that the slopes of the plan lines, given by $(1-t)w$, exceed the slope of the wage income line. But the marginal tax rate for the preferred plan, plan R, is substantially more negative.

Surely Q would not be A's optimal plan in this cost family. He could establish plans that would lead P to make responses that lay much farther out along KK. The respective plan lines would be tangent to P's indifference curves running through the relevant points on KK. In general, we would expect the slopes of the intersecting indifference curves to become steeper the farther out we move on KK.[10] This means the tangent plan lines will have steeper slopes (more negative marginal tax rates) and more negative vertical intercepts (more negative base stipends).

Following this argument to its logical extreme, we would expect A's optimal plan in this cost family to have s and t equal to minus infinity. Before these parameters reached the infinity level, however, P would be working so hard that A would value negatively any additional hours worked. The likely result is that A's optimal plan will have s and t take on substantially, but not ridiculously, negative values.[11]

This within-family optimality property will hold no matter what numerical value is assigned to c^*. This implies that the optimum optimorum plan will also have a substantially negative s and t. Otherwise,

[10] A movement out along KK gives P more of one valued variable (after-transfer income) and less of another (leisure). We would expect the marginal rate of substitution of the first for the second, as indicated by the slope of P's indifference curve through the relevant point on KK, to be increasing. Even short-run exceptions to this expectation can be ruled out, if 1) P's indifference curves are strictly convex, if 2) leisure is not an inferior good for P, and if 3) all other goods taken together (what is in effect P's after-transfer income) are not inferior for him. Conditions 2) and 3) insure, respectively, that upward and rightward movements across P's indifference map will never lead to indifference curves of decreasing steepness. The three conditions together insure that northeasterly movements will lead to curves of increasing steepness.

[11] In a more extensive version of this paper, I solved A's optimization problem explicitly for the utility functions given in equation (4) below. P's reaction function turns out to be $h=1500-s/2w(1-t)$; he will work less as s or t increases. Given that A's utility function values increases in h positively, for all values of h, we get the extra result that A's optimal transfer plan has s and t equal to minus infinity.

there would be a plan within its own cost family to which it would be inferior.

To find his global optimal plan, A must systematically vary the value of c^*, each time finding the optimal plan within the relevant cost family. He then makes a comparison among these plans to find the one that yields him the highest utility. This will be his optimum optimorum plan.

The fact that A's optimal plan has a negative marginal tax rate should not be surprising.[12] We would come to that conclusion by extending the argument relating to the advantages of NIT plans over income maintenance plans. A negative t encourages P to earn more on his own. His resulting higher income offers a greater externality to A; A's bonus-granting procedure is similar to the one used by foundations that give matching grants to support universities and at the same time to increase incentives for alumni contributions.

Results with the Optimal Plan—Parallels with Other Strategic Situations

To determine the optimal assistance program to the poor, I utilized the framework of a two-move, non-zero sum game. In this case, A selected a plan that, when it takes account of P's maximizing responses, leads to the best outcome for A. In this context, I showed that a negative income tax plan with fractional marginal tax rates will have an advantage over its limiting form, the income maintenance plan, a plan with a 100 percent marginal tax rate. Plans with negative marginal tax rates were found to be even more attractive from A's standpoint.

Unfortunately, the plan that is optimal for A in my game-theoretic formulation is in no way optimal for P and A together; it is not a Pareto optimum. The representative citizen and the representative poor man can each take actions that convey positive externalities to the other; they are in a situation of reciprocal externalities. The citizen can increase the transfer he gives to the poor man. The poor man can work longer hours and earn more income, thus indirectly increasing the utility of the citizen. But P acts as a self-interested maximizer and A is only interested in those aspects of P's well-being that enter directly into his own utility function. What results, of course, is an outcome inferior to a cooperative solution.

Situations of reciprocal externalities have received much study in the economics literature. Most frequently the point of departure is the outcome with simultaneous independent adjustment. In this context, Otto Davis and Andrew Whinston discuss two firms each of which conveys production externalities to the other, and James Buchanan and Gordon Tullock consider a community whose members can immunize themselves and thus convey an externality to each other by decreasing the likelihood of exposure to a disease. The simultaneous-independent-adjustment outcome, with each party maximizing in his own self-interest, will be inefficient. The equilibrium has the abstract characteristics of the inefficient, but jointly-dominated, outcome of the well-known prisoners' dilemma. (See R. Duncan Luce and Howard Raiffa, pp. 95–102, and Buchanan.)

In the dynamic aspect of their strategic interaction, Stackelberg's model of leadership duopoly (see William Fellner, pp. 98–119) and A. L. Bowley's model of bilateral monopoly closely resemble the model presented here. There are two key parallels among the three models: One player moves first in each of them. He is the first pick his quantity in the duopoly model. He is the seller who starts by setting the price in the bilateral monopoly model. He is the representative citizen

[12] Jonathan Kesselman also discussed the attractiveness of a negative marginal tax rate for income transfer plans.

who establishes the assistance plan in the income transfer model. Second, the first mover is assumed to know the reaction function of the player who moves second. In moving first he acts strategically and maximizes against the known reaction. The outcome that is achieved in each of these models is dominated by a cooperative solution. Cooperation in these three models would require that their first mover respectively market a smaller quantity, set a lower price, and establish a more generous transfer plan. The second mover would agree not to maximize in a self-interested way against the choice of the first mover. He would market a smaller quantity in the duopoly model, buy a greater quantity in the bilateral monopoly model, and work longer hours in the income transfer model, than he would if he solely considered his own welfare.

III. A Pareto Optimal Transfer Plan

Fortunately, if Pareto optimality is our goal, our analysis has not led us far astray. Plans with substantially negative marginal tax rates can lead to Pareto optima even though P responds as a self-interested maximizer. Except in very perverse cases, plans with positive marginal tax rates will not result in outcomes that are even close to the Pareto frontier.

To see this most clearly, consider hypothetical utility functions for A and for P. They are[13]

(4) $\quad {}_AU = A(y, h, c) = y^{4/10}h^{1/10}(3000-c)^{5/10},$

and

$$ {}_PU = P(y, h) = y^{1/2}(3000-h)^{1/2} $$

To simplify further discussion, I assume that the wage rate, w, is equal to 1.

Figure 3 presents an indifference curve for P and one for A. These curves are derived from the utility functions specified above. From (3), given w, c is uniquely determined once y and h are known. This enables us to represent an indifference curve for A's three-argument utility function in a two-dimensional figure.

If A establishes plan X, P will search along its plan line to find his point of maximum utility. P's maximizing response will be at point x, where the plan line is tangent to his indifference curve.[14] Point x has another important property. It is a point where one of A's indifference curves is tangent to that of P. This means that it is a Pareto optimal point. Thus, plan X leads P to make a response at a Pareto optimal point. It is a Pareto optimal plan.

Plan X has a negative base stipend, $s = -\$5119.37$, which is indicated by the negative intercept of its plan line with the vertical axis. Its negative marginal tax rate, $t = -2.605$, is shown by the fact that the slope of the plan line is greater than the wage rate (in this case 1).

Each Pareto optimal point, each of the points that lies along the contract curve, can be achieved if the linear plan is established tangent to the two indifference curves at their point of tangency. The linear transfer plans that lead to Pareto optimal points, what we might call the Pareto plans, perform a function similar

[13] The parameter assigned the value 3,000 in the utility function for A can be increased (decreased) to make cost a less (more) important consideration for A. This particular utility function for P has a unitary elasticity of substitution. Changing the marginal tax rate will affect hours worked only if there is a non-zero stipend. For positive stipends, lowering the marginal tax rate (an effective wage increase) will increase hours worked and conversely.

[14] P's reaction function is easily derived. He maximizes ${}_PU$. For positive ${}_PU$, he can equally well maximize $({}_PU)^2$.

(a-1) $\quad ({}_PU)^2 = 3000s - sh + 3000hw(1-t) - h^2w(1-t)$

Setting the derivative with respect to h equal to zero, we find

(a-2) $\quad -s + 3000w(1-t) - 2hw(1-t) = 0$

Thus,

(a-3) $$h = 1500 - \frac{s}{2w(1-t)}$$

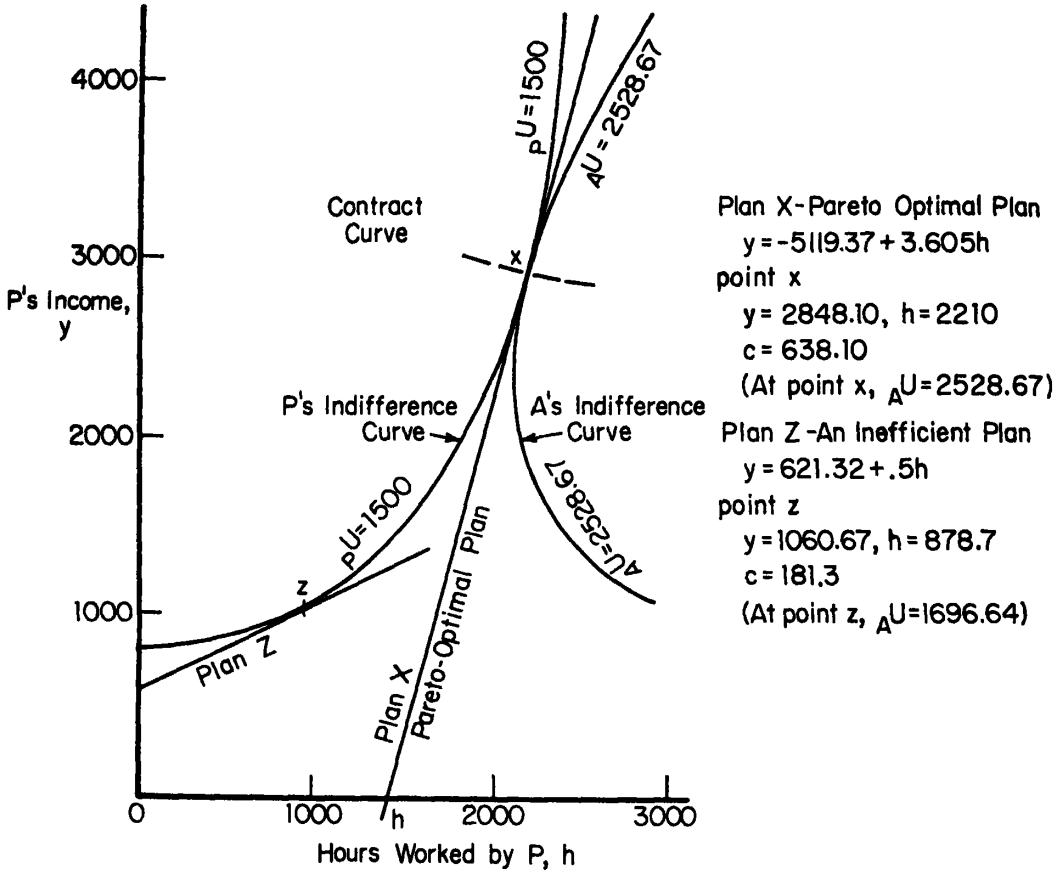


FIGURE 3. ONE PARETO OPTIMAL PLAN

to that of equilibrium price lines in competitive markets. In the case of transfer plans, however, there is no meaningful equivalent to the point of initial endowments.

Of course it would be possible to determine the Pareto optimal points directly, using conventional Lagrangean techniques. Each Pareto optimal point maximizes the sum $P(y, h)+\lambda A(y, h, c)$ for some positive value of λ. Here c is given as function of h and y. For each particular value of λ we can employ conventional maximization techniques to find the y and h that maximize the weighted sum of utilities. By varying the value of λ, we can trace out the locus of efficient trade points, the contract curve.

It is reassuring to observe that simple transfer plans can achieve the same desirable outcomes that could be negotiated directly between A and P. In our society at least, even though the outcomes with the two schemes were the same, it would seem far preferable to establish impersonal pricing schemes that enable individuals to choose how many hours they wish to work, rather than having negotiated outcomes that specify in advance the number of hours to be worked.

Pareto optimal plans have some interesting properties we should note. It is impor-

tant to understand why there is no need to worry with such plans about the externality that *P*'s work decision conveys to *A*. The function of such plans is to externalize (what James Buchanan and William Stubblebine have labeled) Pareto-relevant externalities. The externalization procedure insures that at the equilibrium point *P* and *A* will have the same trade off rate between the two valued variables, *y* and *h*. This equality in rates guarantees that the equilibrium will be efficient.[15]

The conjunction of three facts implies that linear Pareto optimal plans will have negative marginal tax rates whenever *A*'s and *P*'s utility functions have partials with the signs specified above (whenever their indifference curves have the assumed conventional shape). First, *P*'s indifference curves have positive slope. Therefore, with regard to Pareto optimality considerations, that is with regard to possible points of tangency, we need only consider those portions of *A*'s curves that have positive slope as well. Second, *A*'s indifference curves can never have a section with positive slope that does not exceed the wage rate. Otherwise, northeasterly movements along that section of the curve would indicate more hours worked, more income to *P*, at less cost to *A* (at the same cost if the slope just equals the wage rate). This would represent an improvement for *A* on all three valued variables. It would not be a curve of indifference; *A* would not be indifferent between the old point and the "improved" point. Third, a line that is tangent to an indifference curve of *A* and an indifference curve of *P* at their point of tangency will have the same slope that the indifference curves have at that point. The geometric implication of these three facts is that linear Pareto optimal plans will be represented by lines with positive slopes in excess of the wage rate. With such a plan *P* will be able to work an additional hour and receive extra income in excess of his hourly wage. The marginal tax rate on his earnings will be negative.

Plans with positive marginal tax rates are not only not optimal, they are substantially inferior to the Pareto optimal plans, all of which have negative marginal tax rates. Plan *Z* in Figure 3 is the plan with a marginal tax rate of .5 that gets *P* to the same indifference curve he would reach with plan *X*, a Pareto optimal plan.[16] The base stipend of plan *Z* is $621.32. Under plan *Z*, *P* will choose to work the number of hours and receive the income indicated by point *z*. The geometry makes it evident that for *A*, point *z* is far inferior to point *x*.[17]

Earlier we found that the best plan for the representative citizen will have a negative marginal tax rate. We can conclude from this section that this result still holds if the recipient's welfare, as well as that of the representative citizen, is considered.

IV. The Model in Application

My formulation has overlooked the significant fact that the same employment opportunities are not available to all people. A severely handicapped individual, a member of the hard core unemployed, or a person who receives a very meager wage in accord with his marginal product will derive little benefit from an assistance

[15] For example, we might encounter an externality-based inefficiency if my neighbor and I disagreed on the desirability of his spending $1,000 annually to keep his house freshly painted. The efficient arrangement would be for me to compensate him at the margin so that his after-compensation trade off rate was the same as mine.

[16] There is no plan with a .5 marginal tax rate which leads to a point on the indifference curve of *A* which runs through point *x*.

[17] A simple calculation reveals the magnitude of the inferiority. If we hold *y*, *h*, and *c* to their values at *z*, it would take a lump sum payment of $3,672 (which *A* could apply against cost) to get *A* to his indifference curve which runs through point *x*. That is, $A(1060.67, 878.7, 3000-(181.3-3672))=2528.67$, the value of $_AU$ at point *x*.

plan that channels all of its positive transfer through a substantial bonus on wage income.

In a more extensive version of this paper, I have shown when the opportunities for poor men differ, the optimal income transfer program will offer more than one transfer plan. For those with little or no opportunity to earn, a plan with a sizable positive stipend will be used to provide assistance. Those who could find productive work will receive assistance under a bonus incentive plan of the type described in this paper.

REFERENCES

A. L. Bowley, "On Bilateral Monopoly," *Econ. J.*, 1928, *38*, 651–59.

J. M. Buchanan, "Cooperation and Conflict in Public-Goods Interaction," *Western Econ. J.*, Mar. 1967, *5*, 109–21.

——— and W. C. Stubblebine, "Externality," *Economica*, Nov. 1962, *29*, 371–84.

——— and G. Tullock, "Public and Private Interaction under Reciprocal Externality," in J. Margolis, ed., *The Public Economy of Urban Communities*, Washington 1965, 52–73.

O. A. Davis and A. Whinston, "Externalities, Welfare, and the Theory of Games," *J. Polit. Econ.*, June 1962, *70*, 241–62.

P. A. Diamond, "Negative Taxes and the Poverty Problem—A Review Article," *Nat. Tax J.*, Sept. 1968, *21*, 288–303.

W. Fellner, *Competition Among the Few: Oligopoly and Similar Market Structures*, New York 1949.

M. Friedman, *Capitalism and Freedom*, Chicago 1962.

C. Green, *Negative Taxes and the Poverty Problem*, Washington 1967.

J. Kesselman, "Labor-Supply Effects of Income, Income-Work, and Wage Subsidies," *J. Hum. Resources*, summer 1969, *4*, 275–92.

R. D. Luce and H. Raiffa, *Games and Decisions*, New York 1957.

M. Olson, Jr., *The Logic of Collective Action: Public Goods and the Theory of Groups*, New York 1968.

P. A. Samuelson, "The Pure Theory of Public Expenditure," *Rev. Econ. Statist.*, Nov. 1954, *36*, 387–89.

R. Theobald, *The Guaranteed Income*, New York 1966.

J. Tobin, "The Case for an Income Guarantee," *Publ. Interest*, summer 1966, No. 4, 31–41.

——— et al., "Is a Negative Income Tax Practical?", *Yale Law J.*, Nov. 1967, *77*, 1–27.

J. Vadakin, "A Critique of the Guaranteed Annual Income," *Publ. Interest*, spring 1968, No. 11, 53–66.

R. Zeckhauser, "Optimal Mechanisms for Income Transfer," RAND P-3878, June 1968, and Harvard Inst. Econ. Res. disc. paper 40, Aug. 1968.

——— and P. Schuck, "An Alternative to the Nixon Income Maintenance Plan: A Solution to the Problem of Work Incentives," *Publ. Interest*, spring 1970, No. 19, 120–130.

Poverty Amid Plenty: The American Paradox, *The Report of the President's Commission on Income Maintenance Programs*, Ben Heineman, chairman, Nov. 12, 1969, Washington.

The Optimal Quantity of Money, Bonds, Commodity Inventories, and Capital

By Edgar L. Feige and Michael Parkin*

Economists have long concerned themselves with the problem of finding a satisfactory means of incorporating money into the theory of economic behavior. That this problem is still far from being solved is apparent in the surprisingly large discrepancy between the rhetoric of the monetary economist and his analysis. The hallmark of money as extolled in most textbooks is its function as a medium of exchange—yet we have very little in the way of an explicit theory of exchange. Money, that "unique" asset, is generally treated as just another consumer durable good which yields a flow of ill-defined nonobservable services offering the individual "convenience, security and liquidity." The traditional neoclassical approach, underlying much of the important work of several authors (see articles by Milton Friedman, Harry Johnson, Don Patinkin, and Paul Samuelson), introduces real cash balances directly into the utility function, and pays lip service to the costs of exchanges between money and goods although these costs are not made explicit in the analysis. However, once money is treated as an argument of the utility function, many of the interesting questions concerning what money does, and why people hold it—are essentially stifled, since the model does not specify the explicit role money plays as a medium of exchange.

An alternative approach dealing more directly with the transaction aspects of money is the Baumol-Tobin inventory theoretic framework (see articles by William Baumol and James Tobin). The Baumol-Tobin analysis is framed in terms of the explicit costs of making transactions in the money-bond market and thus represents a beginning toward a theory of exchange behavior or its analytic equivalent of optimal inventory theory. Although the theory focuses explicit attention on the costs of exchanging bonds and money, it ignores the costs of exchanging money for commodities and thus neglects the holdings of commodity inventories. Baumol and Tobin leave the utility function unspecified and as such treat total consumption expenditures as exogenously determined. Money is left to enter a budget constraint which involves minimum cost management of bond and money inventories.

Recently, a number of economists have refocused their attention on the rather fundamental problem of formulating a useful theory of the role of money in individual behavior (see articles by Robert Clower, Feige and Donald Nichols, Friedman (1969), Johnson, David Laidler, Alvin Marty, Samuelson). The impetus for this work has been the formulation of a provocative problem in monetary theory; namely, what is the optimal quantity of money and

* Associate professor of economics, University of Wisconsin; professor of economics, University of Manchester, respectively. The authors wish to acknowledge helpful discussions with Laurits Christensen, David Laidler, Donald Nichols, Kenneth Smith, and the excellent assistance of Yung San Lee and Rodney J. Barrett. The Federal Deposit Insurance Corporation, the University of Wisconsin Graduate School, and the Social Science Research Council (U.K.), provided financial support for the project. An earlier version of this paper was presented in March 1969 at the University of Essex.

how can individuals be induced to hold the optimal quantity? Furthermore, in what sense is this an optimal quantity of money and what are the consequences of its achievement?

This paper has two aims. First, to extend the Baumol-Tobin inventory theoretic approach to incorporate costly transactions between money and commodities. Second, using this broadened inventory theoretic framework, to examine the question of what is the optimal quantity of money.

Our extension of the Baumol-Tobin inventory model denies that money yields any direct services which enter as arguments of the utility function. Money is simply used to facilitate transactions in the bond and commodity markets; and, due to its technological property of being an efficient medium of exchange, its use allows the individual to economize on other real resources which would otherwise be expended in the process of exchange. In this view, the quantity of money enters into the individual's budget constraint, and his inventory management decisions affect the amount of real income available for consumption of other goods and services.

Our aim is to examine this alternative specification of how money affects individual behavior. We wish to determine whether or not the inventory theoretic approach gives rise to conclusions concerning the optimal quantity of money, which are similar to those offered by the neoclassical approach.

It has been claimed by numerous authors that, to the extent that the marginal social cost of increasing the supply of money is zero, individuals should hold money to the point of satiety (see articles by Johnson and Samuelson)—that is, to the point where the marginal utility of holding money is zero. Since money does not bear interest, the opportunity cost to the individual holding part of his wealth in the form of cash balances is the foregone interest on those cash balances. Individuals will thus be induced to economize on cash balance holdings, and the resulting quantity of real cash balances in the society will be less than the Pareto optimal quantity. It is precisely these attempts to economize on real cash balances that give rise to the using up of real resources for transactions purposes and thus to a welfare loss. From society's point of view it is virtually costless to increase the nominal supply of money; but the *real* quantity in existence is in fact determined by the demanders of money. Thus, if the government is to induce people to hold larger real cash balances it must provide an institutional inducement which may or may not be costless. One way in which the community can be persuaded to hold more real cash balances is to pay a rate of interest on money. Alternatively (and equivalent analytically) the government can induce a continuous deflation such that there is an implicit rate of return in money holdings due to the expected increase in their value.

The following analysis concerns itself entirely with the use of paying interest on money as a means of encouraging the community as a whole to hold larger real cash balances. Thus, the "supply of *real* cash balances" can be increased only by affecting the "quantity demanded" of real cash balances, and that is accomplished by paying interest on money. The institutional arrangement which provides for interest payments is unlikely to be costless. If it were costless we would expect the optimal quantity of real cash balances to be the satiety level of real cash balances, i.e., that level of real cash balances which makes the marginal utility of money balances go to zero. This is perfectly consistent with Pareto-efficiency: if the marginal

social cost of producing real balances is zero, then the real balances are a free good and marginal utility should be equated with marginal social cost.

The foregoing arguments are generally well known, but there is little evidence to suggest that their implications are well understood. For example, what other consequences flow from real cash balances being a free good? We know that the quantity of real cash balances will increase, but we know little explicitly about the other effects of such a policy. Samuelson has speculated that a satiety policy would result in "fewer trips to the bank and to the broker, . . . smaller printing costs and other costs of transactions" (p. 10). Such explanations, however, are entirely *ex post* rationalizations and certainly cannot be deduced from the current specification of the theory. A major difficulty lies in the manner in which money is typically introduced into the theory of economic behavior. Money is either treated as a consumer durable which yields an unobservable flow of nonpecuniary services or (analytically equivalent and equally unilluminating for the problem at hand) incorporated into the production function as just another input.

Both formulations have contributed something to our general understanding. The consumer theory approach has provided the underlying rationalization for most of the empirical studies of the demand for money—arguing that, as a first approximation, one can regard the flow of nonpecuniary services as proportional to the stock of real cash balances which are observable and thus bring standard consumer theory to bear on the question of the demand for money. When money is included in the utility function for individuals, it can be demonstrated that individuals will hold a less-than-optimal quantity of money. However, the sources of the gains from inducing individuals to hold an optimal quantity of money cannot be directly deduced from such a weakly specified analysis.

Our extension of the inventory theory framework seeks to avoid this problem by explicitly examining the effects of a satiety or free money policy on the quantity of real cash balances held as inventories, as well as its effects on the holdings of bonds, commodity inventories, and productive capital. The analysis also highlights how a free money policy will increase some types of transactions while reducing others. Finally, the theory is capable of yielding exact expressions for the net gain in welfare resulting from a free money policy. The analysis abstracts from problems of risk and uncertainty and also works within the framework of the stationary state.

I. An Extended Inventory Theoretic Model of the Demand for Money, Bonds, and Commodity Inventories—The Individual Optimum

Throughout, we assume that the monetary authority controls the *nominal* money stock in such a way as to maintain a stable price level. The monetary authority is permitted to influence the demand for *real* money balances by paying interest on its monetary obligations. When it does so, or when it varies that rate, it is assumed to simultaneously vary the *nominal* money supply in such a way that the price level is undisturbed. It is worth noting that this assumption is exactly the opposite of that used in much of the growth literature, which treats the rate of inflation (deflation) as a policy parameter and interest rates as market determined. We, instead, treat the rate of interest on the government's monetary obligations as the policy parameter. Thus our analysis abstracts entirely from the dynamics of anticipated and unanticipated changes in the price

level, and allows us to treat real cash balances and nominal cash balances interchangeably. A similar correspondence holds between real bond holdings and nominal bond holdings. We could, of course, achieve analytically equivalent results by letting the authorities reduce the nominal money stock at a constant rate, thereby generating a constant rate of (fully anticipated) deflation in the price level.

In addition to assuming fixed prices, we assume a fixed single period of analysis and stationarity.

The economy consists of representative family units which maximize the utility function

$$U = U(pq) \tag{1}$$

where U is the utility index, and pq is the dollar volume of commodities consumed.

The utility function is maximized subject to the constraint

$$0 = Y + \pi - pq - T \tag{2}$$

where

Y = labor income in current dollars
π = net profit from inventory management
pq = expenditure on commodities
T = taxes

Taxes, which are viewed as exogenous by the individual family unit comprise a neutral transfer component and a component involving the real resource costs of operating an interest payments mechanism on balances. Thus

$$T = t + h(r_m), \tag{3}$$

where

t = transfer component
$h(r_m)$ = real resource component
$h(r_m) = 0$ for $r_m = 0$
$h(r_m) = k$ for $r_m > 0$

It is the variable π which represents the novelty of the inventory theoretic approach, and through which we introduce explicitly the role of money as it influences individual behavior. We define π as the profit from inventory management of capital, bonds, money and commodities, net of transaction costs and inventory carrying costs. Specifically,

$$\begin{aligned} \pi = {} & (r_k - \alpha_k)p\overline{K} + (r_b - \alpha_b)\overline{B} \\ & + (r_m - \alpha_m)\overline{M} - \alpha_q\overline{pQ} \\ & - \beta_b n - \beta_q m, \end{aligned} \tag{4}$$

where

$p\overline{K}$ = average stock of capital in current dollars
$\overline{B}$ = average inventory of bonds in current dollars
$\overline{M}$ = average inventory of cash balances
$\overline{pQ}$ = average inventory of commodities in current dollars
r_k = rate of return on capital
r_b = rate of return on bonds
r_m = rate of return on cash balances
α_k = the cost per dollar of carrying capital inventories
α_b = the cost per dollar of carrying bond inventories
α_m = the cost per dollar of carrying cash balances
α_q = the cost per dollar of carrying commodity inventories
n = number of bond market transactions per period
m = number of commodity market transactions per period
β_b = the cost per bond market transaction
β_q = the cost per commodity market transaction.

Most of the variables in the definition of π are familiar; however, the interpretation of the α's and β's require some further elaboration. The α's are inventory carrying cost parameters which reflect the real re-

source costs of carrying inventories of money, bonds, commodities, and physical capital. They might include such items as the costs of insurance, safety deposit boxes, refrigeration services, and other costs incurred in the storage of inventories. The β's reflect the costs associated with particular transactions such as time, brokerage fees, and delivery charges.

We assume that the individual receives income at the beginning of the period and that he spreads out his total consumption, pq, evenly throughout the period. In order to consume, the individual must maintain some stock of commodity inventories; and we assume that the individual runs down his inventory of commodities at a constant rate. The individual acquires commodity inventories by making m equally spaced trips to the store. Since the total value of consumption for the period is pq, the amount spent per trip to the store is simply pq/m and on average, commodity inventories for the period are

$$p\overline{Q} = \frac{pq}{2m} \tag{5}$$

We assume that all receipts come in the form of money. The individual makes, during the period, n equally spaced trips to the broker. On the first trip he buys bonds. On the subsequent $n-1$ trips he sells bonds. On the first trip he leaves himself with pq/n of cash, buying bonds of $(n-1/n)pq$. On the subsequent $n-1$ trips he cashes pq/n of bonds. Having received pq/n in cash (retained pq/n in the case of the first trip), the individual immediately spends pq/m on commodities. Thus he starts out after a trip to the broker with $pq/n-pq/m$ in cash. Since he runs down this cash with evenly spaced trips to the store, his average cash balance is

$$\overline{M} = \frac{pq}{2n} - \frac{pq}{2m} \tag{6}$$

Finally, average holdings of bond inventories for the period are simply $(n-1/2n)pq$, so that,

$$\overline{B} = \frac{pq}{2} - \frac{pq}{2n} \qquad n \geq 1 \tag{7}$$

It should be noted that m and n are integers and that m is a multiple of n. The condition that $n=m$ implies that there is perfect synchronization between bond market and commodity market transactions, such that cash balance inventories are zero.

If the broker is regarded as a bank which provides checking services as well as bond market brokerage services, we can give the analysis a somewhat broader interpretation. In this case the individual is assumed to be paid in cash and makes a trip to the banker-broker to deposit cash of pq/n while purchasing bonds of $(n-1/n)pq$. If commodity purchases are paid by check, we consider the marginal cost of checking services to be associated with trips to the store and thus reflected in β_q. In this case β_b is left to reflect the initial cost of establishing a bank account and the brokerage costs of subsequent conversions of bonds into checking deposits. For the special cases where $n=1$, the individual simply deposits his entire income at the banker-broker and purchases no bonds. Thus the average deposit balance inventory becomes $pq/2-pq/2m$ and average bond holdings for the period are zero.

The individual is also subject to a wealth constraint such that his average fixed stock of assets is denoted as

$$\overline{A}^* = p\overline{K} + \overline{B} + \overline{M} + p\overline{Q} \tag{8}$$

Thus,

$$p\overline{K} = \overline{A}^* - \frac{pq}{2} \tag{9}$$

Equation (8) captures the opportunity cost to the individual of adding to cash

balances or commodity inventories, because any such transfer can only be accomplished by sacrificing an equivalent dollar volume of bonds or capital.

The representative family must therefore choose the pq, m, and n that will maximize the utility function subject to his budget and wealth constraints. Substituting (4), (5), (6), (7), and (8) into (2), we form the Lagrangean:[1]

$$
\begin{aligned}
V &= U(pq) \\
&\quad + \lambda\left[Y + (r_k - \alpha_k)\left(\overline{A}^* - \frac{pq}{2}\right)\right. \\
&\quad + (r_b - \alpha_b)\left(\frac{pq}{2} - \frac{pq}{2n}\right) \\
&\quad + (r_m - \alpha_m)\left(\frac{pq}{2n} - \frac{pq}{2m}\right) \\
&\quad \left. - \alpha_q\frac{pq}{2m} - \beta_b n - \beta_q m - pq - T\right]
\end{aligned} \tag{10}
$$

[1] Although inventory theory specifies n and m as integers, we treat the Lagrangean as continuously differentiable. This treatment can result in some minor departures from the minimum cost conditions. We chose to ignore this difficulty and throughout the paper allow n and m to take on non-integer values as determined by the parameters of the model.

Differentiating (10) with respect to pq, n, m, and λ and setting the partial derivatives equal to zero gives the system of equations (11).

In order to compare the results implied by equations (11) with the more familiar Baumol inventory model, we first examine a partial solution which considers pq as predetermined. Under this assumption, equations (11b) and (11c) can be solved for n and m as functions of pq to determine the optimal number of trips to the broker and the store:

$$n^* = \sqrt{\frac{(r_b - \alpha_b - r_m + \alpha_m)}{2\beta_b}pq} \tag{12}$$

$$m^* = \sqrt{\frac{(r_m - \alpha_m + \alpha_q)}{2\beta_q}pq} \tag{13}$$

Assuming a given level of consumption, equations (12) and (13) reveal how interest payments on cash balances affect transactions behavior. Non-zero interest payments on money will reduce trips to the broker but will increase trips to the store. Conversely, as the costs of carrying money balances increase, a greater number of

$$
\begin{aligned}
&\text{(a)}\quad V_{pq} = U'(pq) - \lambda\left[\frac{(r_k - \alpha_k - r_b + \alpha_b)}{2} + \frac{(r_b - \alpha_b - r_m + \alpha_m)}{2n}\right. \\
&\qquad\qquad \left. + \frac{(r_m - \alpha_m + \alpha_q)}{2m} + 1\right] = 0 \\
&\text{(b)}\quad V_n = \lambda\left[\frac{(r_b - \alpha_b - r_m + \alpha_m)}{2n^2}pq - \beta_b\right] = 0 \\
(11)\quad &\text{(c)}\quad V_m = \lambda\left[\frac{(r_m - \alpha_m + \alpha_q)}{2m^2}pq - \beta_q\right] = 0 \\
&\text{(d)}\quad V_\lambda = \left[Y + (r_k - \alpha_k)\left(\overline{A}^* - \frac{pq}{2}\right) + (r_b - \alpha_b)\left(\frac{pq}{2} - \frac{pq}{2n}\right)\right. \\
&\qquad\qquad \left. + (r_m - \alpha_m)\left(\frac{pq}{2n} - \frac{pq}{2m}\right) - \alpha_q\frac{pq}{2m} - \beta_b n - \beta_q m - pq - T\right] = 0
\end{aligned}
$$

bond market transactions will be undertaken and a smaller number of commodity market transactions will be made. Thirdly, increased transactions costs will reduce the number of transactions in the respective markets.

Substituting (12) and (13) into (5), (6), and (7) yields partial demand functions for average cash balances, average commodity inventories and average bond holdings. These are:

$$(14)\quad (\overline{M})^* = \sqrt{\frac{\beta_b pq}{2(r_b - \alpha_b - r_m + \alpha_m)}} - \sqrt{\frac{\beta_q pq}{2(r_m - \alpha_m + \alpha_q)}}$$

$$(15)\quad (p\overline{Q})^* = \sqrt{\frac{\beta_q pq}{2(r_m - \alpha_m + \alpha_q)}}$$

$$(16)\quad (\overline{B})^* = \frac{pq}{2} - \sqrt{\frac{\beta_b pq}{2(r_b - \alpha_b - r_m + \alpha_m)}}$$

Equations (12) to (16) suggest that a change in r_b directly affects the number of trips to the broker but does not affect the number of trips to the store. Hence, bond and money inventories are switched but commodity inventories are unaffected.[2] That is, a change in r_b directly affects only the margin of substitution between bonds and money. A change in r_m affects the margin of substitution between money and bonds as well as between money and commodity inventories.

Consider the case where both r_b and r_m rise by the same amount leaving invariant $(r_b - r_m)$. Traditionally, $(r_b - r_m)$ has been treated as the opportunity cost of holding real balances; an unchanged $(r_b - r_m)$ would leave invariant the stock of real balances held. This result remains intact in the Baumol model even when explicit interest is paid on money. But when the model is extended to the margin of substitution between money and commodities, an increase in r_b and r_m which leaves the difference unchanged, alters the stock of real balances held. This result arises from the fact that r_m affects the margin of substitution between bonds and money on the one hand, and between money and commodities on the other, whereas r_b affects only the margin between bonds and money. This implies that the demand function for real balances is not a unique function of the difference between r_b and r_m but instead depends on their levels. Of course, what has to be noticed here is that in an n asset world, there are $n-1$ margins of substitution and only if *all* rates move by the same absolute amount will demands be invariant to the rate levels. In our model, we can interpret α_q (the carrying cost of commodity inventories) as the (negative) rate of return on commodity inventories. An increase in r_b and r_m with an equal change (decrease) in α_q will in fact leave all quantities demanded unchanged.

The demand function for cash balances (14) reduces to the familiar Baumol result given Baumol's more restrictive assumptions.

In Baumol's model, pq is predetermined; there are no costs associated with the purchase of commodities; no interest payments on real cash balances; and no costs of carrying money or bond inventories. Thus, $\beta_q = r_m = \alpha_m = \alpha_b = 0$, and (14) reduces to the Baumol demand function,

$$(17)\quad (\overline{M})^* = \sqrt{\frac{\beta_b pq}{2r_b}} = \frac{1}{2}\sqrt{\frac{2\beta_b pq}{r_b}}$$

In order to determine the individual's final equilibrium values of n, m, and pq, we drop the assumption that pq is pre-

[2] Note this is only true when pq is predetermined. When pq is endogeneously determined, a change in r_b will have an indirect effect on commodity inventories due to the induced change in total income.

determined, and simultaneously solve the first-order conditions of equation (11) subject to two additional constraints that must be satisfied for the society as a whole. The first constraint simply states that taxes must equal the interest payments on bonds and money plus the real resource costs of instituting an interest payments mechanism on money. Thus:

$$(18)\qquad T = r_b\overline{B} + r_m\overline{M} + h(r_m),$$

where

$$h(r_m) = 0 \qquad \text{for } r_m = 0$$
$$h(r_m) = k \qquad \text{for } r_m > 0$$

and the tax transfer scheme is assumed to have neutral allocation consequences.

The final constraint derived from the stationary assumption, states that society's stock of inventories of physical capital plus the stock of commodity inventories is constant.[3] Thus:

$$(19)\qquad \overline{W}^* = p\overline{K} + p\overline{Q}$$

or

$$(19')\qquad p\overline{K} = \overline{W}^* - \frac{pq}{2m}$$

[3] It should be noted that $\overline{W}^*$ is not the same thing as accounting net worth. If $r_m=0$, accounting net worth is $\overline{W}^*+\overline{M}$. With $0<r_m<r_k$, $\overline{W}^*<$ accounting net worth $<\overline{W}^*+\overline{M}$.

$$(20)\qquad (pq)^* = \frac{X \pm \sqrt{X^2 - 4A^2B^2}}{2B^2}$$

where

$$\begin{aligned} X = {} & (2+\alpha_b)[Y + (r_k - \alpha_k)\overline{W}^* - h(r_m)] \\ & + \left[\frac{(\alpha_m - \alpha_b)^2}{2(r_b - \alpha_b - r_m + \alpha_m)} + \frac{1}{2}(r_b - r_m) + \frac{3}{2}(\alpha_m - \alpha_b)\right]\beta_b \\ & + \left[\frac{(r_k - \alpha_k - \alpha_m + \alpha_q)^2}{2(r_m - \alpha_m + \alpha_q)} + \frac{1}{2}(2r_k - 2\alpha_k + r_m - 3\alpha_m + 3\alpha_q)\right]\beta_q \\ & + \left[\sqrt{\frac{(\alpha_m - \alpha_b)^2(r_k - \alpha_k - \alpha_m + \alpha_q)^2}{(r_m - \alpha_m + \alpha_q)(r_b - \alpha_b - r_m + \alpha_m)}} + \sqrt{\frac{(\alpha_m - \alpha_b)^2(r_m - \alpha_m + \alpha_q)}{(r_b - \alpha_b - r_m + \alpha_m)}}\right. \\ & + \sqrt{\frac{(r_k - \alpha_k - \alpha_m + \alpha_q)^2(r_b - \alpha_b - r_m + \alpha_m)}{(r_m - \alpha_m + \alpha_q)}} \\ & \left. + \sqrt{(r_m - \alpha_m + \alpha_q)(r_b - \alpha_b - r_m + \alpha_m)}\right]\sqrt{\beta_b\beta_q} \end{aligned}$$

$$A = Y + (r_k - \alpha_k)\overline{W}^* - h(r_m)$$

$$B = \tfrac{1}{2}(2 + \alpha_b)$$

$$(21)\qquad n^* = \sqrt{\frac{(r_b - \alpha_b - r_m + \alpha_m)}{2\beta_b}pq,^*}$$

$$(22)\qquad m^* = \sqrt{\frac{(r_m - \alpha_m + \alpha_q)}{2\beta_q}pq^*}$$

Simultaneous solution of (11), (18), and (19) yields quadratic solution equations for determining the equilibrium values of pq, n, and m.[4] These are shown in equations (20)–(22).

Substituting equation (20) into equations (14), (15), and (16) reveals that the demand for average cash balances, commodity inventories, and bonds, depends upon disposable human and nonhuman income—as well as interest rates, transactions costs, and inventory carrying costs.

II. The Social Optimum

The foregoing analysis has determined the optimal equilibrium values of consumption, number of transactions, and inventory holdings, for the individuals in society. In order to determine the corresponding optimal values for society as a whole, it is necessary to examine the sources and uses of resources for the entire society. Given the stationarity constraint and assuming a total population of N families, the total nonhuman wealth for the society is simply:

$$N\overline{W}^* = N(\overline{pK} + \overline{pQ}), \tag{23}$$

and the total income available to society is

$$N[Y + r_k(\overline{W}^* - \overline{pQ})] \tag{24}$$

Thus, for society as a whole, the opportunity cost of holding commodity inventories is the income foregone by not utilizing these inventories as productive capital. The society must allocate its resource flow among the following uses:

(a) holding money at a cost of

$$N\left[\alpha_m\left(\frac{pq}{2n} - \frac{pq}{2m}\right)\right]$$

(b) holding commodity inventories at a cost of

$$N\left[\alpha_q\left(\frac{pq}{2m}\right)\right]$$

(c) holding bond inventories at a cost of

$$N\left[\alpha_b\left(\frac{pq}{2} - \frac{pq}{2n}\right)\right]$$

(d) holding capital inventories at a cost of

$$N\left[\alpha_k\left(W^* - \frac{pq}{2m}\right)\right]$$

(e) transacting in bond markets at a cost of $N\ [\beta_b n]$

(f) transacting in commodity markets at a cost of $N[\beta_q m]$

(g) operating a scheme of paying interest on money of $N[h(r_m)]$, and

(h) consuming, $N[pq]$

The social constraint therefore becomes:

$$0 = N\left[Y + r_k\left(\overline{W}^* - \frac{pq}{2m}\right) - \alpha_m\left(\frac{pq}{2n} - \frac{pq}{2m}\right) - \alpha_q\left(\frac{pq}{2m}\right) - \alpha_b\left(\frac{pq}{2} - \frac{pq}{2n}\right) - \alpha_k\left(\overline{W}^* - \frac{pq}{2m}\right) - \beta_b n - \beta_q m - h(r_m) - pq\right] \tag{25}$$

The solution to the social problem is ob-

[4] It will be noticed that we can solve $(pq)^*$ despite the fact that an unspecified function of $pq - U'(pq)$ appears in the first-order conditions. This is because the first-order conditions form a recursive system and (11a) solves for λ (the Lagrange multiplier) given the values of the other variables from (11b), (11c), and (11d).

tained by maximizing the Lagrangean:

$$(26)\quad V = N\left\{U(pq) + \lambda\left[Y + (r_k - \alpha_k)\cdot\left(W^* - \frac{pq}{2m}\right) - \alpha_m\left(\frac{pq}{2n} - \frac{pq}{2m}\right) - \alpha_b\left(\frac{pq}{2} - \frac{pq}{2n}\right) - \alpha_q\frac{pq}{2m} - \beta_b n - \beta_q m - h(r_m) - pq\right]\right\}$$

Differentiating V and setting the partial derivatives equal to zero yields the system (27).

The first-order conditions (27) can be simultaneously solved to determine the socially optimum level of consumption expenditures, as well as the socially optimum number of transactions. The solution equations are as follows:

$$(28)\quad Npq^* = N\left[\frac{Z \pm \sqrt{Z^2 - 4A^2B^2}}{2B^2}\right],$$

where

$$Z = [Y + (r_k - \alpha_k)\overline{W}^* - h(r_m)]\cdot(2 + \alpha_b) + 2(\alpha_m - \alpha_b)\beta_b + 2(r_k - \alpha_k - \alpha_m + \alpha_q)\beta_q + 4\sqrt{(\alpha_m - \alpha_b)(r_k - \alpha_k - \alpha_m + \alpha_q)}\cdot\sqrt{\beta_b\beta_q}$$

$$A = Y + (r_k - \alpha_k)\overline{W}^* - h(r_m)$$

$$B = \frac{(2 + \alpha_b)}{2},$$

$$(29)\quad Nn^* = N\sqrt{\frac{(\alpha_m - \alpha_b)}{2\beta_b}pq,^*}$$

and

$$(30)\quad Nm^* = N\sqrt{\frac{(r_k - \alpha_k - \alpha_m + \alpha_q)}{2\beta_q}pq^*}$$

In order to compare the socially optimum number of transactions and inventories with the individual optimum values, we construct Table 1, which summarizes the solutions for the social and individual optimum values as functions of pq.

Table 1 reveals that the only way that individuals can be induced to undertake the socially optimum number of transactions, and (equivalent analytically) hold the socially optimum stocks of inventories, is to pay interest on cash balances and bonds equal to the *net* rate of return on capital. Similarly, comparing equations (20) and (28) reveals that optimum individual consumption will equal optimum social consumption if, and only if,

$$(31)\quad r_m = r_b = r_k - \alpha_k$$

This result caused us some surprise at first. Why is not the optimality condition

$$(31a)\quad r_m - \alpha_m = r_b - \alpha_b = r_k - \alpha_k?$$

$$\text{(a)}\quad V_{pq} = N\left\{U'(pq) - \lambda\left[\frac{\alpha_b}{2} + \frac{(r_k - \alpha_k - \alpha_m + \alpha_q)}{2m} + \frac{(\alpha_m - \alpha_b)}{2n} + 1\right]\right\} = 0$$

$$\text{(b)}\quad V_n = N\lambda\left[\frac{(\alpha_m - \alpha_b)}{2n^2}pq - \beta_b\right] = 0$$

$$(27)\quad \text{(c)}\quad V_m = N\lambda\left[\frac{(r_k - \alpha_k - \alpha_m + \alpha_q)}{2m^2}pq - \beta_q\right] = 0$$

$$\text{(d)}\quad V_\lambda = N\left[Y + (r_k - \alpha_k)\left(\overline{W}^* - \frac{pq}{2m}\right) - \alpha_m\left(\frac{pq}{2n} - \frac{pq}{2m}\right) - \alpha_b\left(\frac{pq}{2} - \frac{pq}{2n}\right) - \alpha_q\frac{pq}{2m} - \beta_b n - \beta_q m - h(r_m) - pq\right] = 0$$

TABLE 1—OPTIMAL TRANSACTIONS AND INVENTORY HOLDINGS

Variable	Individual Optimum	Social Optimum
$(n)^*$	$\sqrt{\frac{(r_b - \alpha_b - r_m + \alpha_m)pq}{2\beta_b}}$	$\sqrt{\frac{(\alpha_m - \alpha_b)pq}{2\beta_b}}$
$(m)^*$	$\sqrt{\frac{(r_m - \alpha_m + \alpha_q)pq}{2\beta_q}}$	$\sqrt{\frac{(r_k - \alpha_k - \alpha_m + \alpha_q)pq}{2\beta_q}}$
$(\bar{M})^*$	$\sqrt{\frac{\beta_b pq}{2(r_b - \alpha_b - r_m + \alpha_m)}} - \sqrt{\frac{\beta_q pq}{2(r_m - \alpha_m + \alpha_q)}}$	$\sqrt{\frac{\beta_b pq}{2(\alpha_m - \alpha_b)}} - \sqrt{\frac{\beta_q pq}{2(r_k - \alpha_k - \alpha_m + \alpha_q)}}$
$(\bar{B})^*$	$\frac{pq}{2} - \sqrt{\frac{\beta_b pq}{2(r_b - \alpha_b - r_m + \alpha_m)}}$	$\frac{pq}{2} - \sqrt{\frac{\beta_b pq}{2(\alpha_m - \alpha_b)}}$
$(\overline{pQ})^*$	$\sqrt{\frac{\beta_q pq}{2(r_m - \alpha_m + \alpha_q)}}$	$\sqrt{\frac{\beta_q pq}{2(r_k - \alpha_k - \alpha_m + \alpha_q}}$
$(p\bar{K})^*$	$\bar{W}^* - \sqrt{\frac{\beta_q pq}{2(r_m - \alpha_m + \alpha_q)}}$	$\bar{W}^* - \sqrt{\frac{\beta_q pq}{2(r_k - \alpha_k - \alpha_m + \alpha_q)}}$

If r_m and r_b were fixed to achieve the equalities specified in (31a) instead of those in (31), the private opportunity cost of the services from real balances of money and bonds would be zero. But α_m and α_b represent real social opportunity costs of holding money and bonds (and hence of receiving the service yield from those holdings). Optimality requires that the private opportunity costs must equal the social costs, hence the private cost of holding money and bonds must be α_m and α_b, respectively. This is achieved by paying interest at r_m and r_b equal to the net return on physical capital and having the individual bear the carrying costs α_m and α_b.

It may still be objected that (31a) must be satisfied because net rates of return must be equalized. In fact, net rates of return are equalized in (31). They are unequal in (31a). The reason for this is that money and bonds both yield returns in the form of the increased consumption made possible by lowering inventory management costs. These returns are additional to the interest income that will arise from holding money and bonds. The optimality conditions (31) suggest that these additional service yields from money and bonds are equal on the margin to α_m and α_b, respectively.

III. Welfare Implications

We can now explicitly determine the sources and possible magnitudes of the net welfare gain resulting from the institution of a payments mechanism on cash balances which satisfies equation (31). To this end we consider two regimes.

In regime I, we assume that bonds yield the net rate of return on capital, but no interest is paid on cash balances. Regime II is similar to regime I except that interest is paid on cash balances equal to the return on bonds and the net return on capital. We assume for the moment that it is costless to introduce an interest-payments mechanism on money. We define the gross welfare gain to society from the institution of a payments mechanism on money as that increase in consumption which can be indefinitely maintained by shifting from regime I to regime II. The consumption level under the two regimes can be obtained directly by solving the quadratic solution equation (20) under the respec-

tive sets of assumptions. To use a numerical illustration, we assume the following values for the parameters of the model:

(32)

$$\begin{aligned} Y &= \$10{,}000.00 & \alpha_m &= .01 \\ W^* &= \$50{,}000.00 & \beta_b &= \$15.00 \\ r_k &= .15 & \beta_q &= .20 \\ \alpha_k &= .09 & \alpha_q &= .06 \\ r_b &= .06 & h(r_m) &= 0 \\ \alpha_b &= .005 & r_m &= 0 \text{ Regime I} \\ N &= 50{,}000{,}000 & r_m &= .06 \text{ Regime II} \end{aligned}$$

Solving (20), given the foregoing values of the parameters, we derive two roots for each regime. The higher root exceeds the budget constraint and therefore only the legitimate lower root is presented in Table 2.

TABLE 2—OPTIMUM CONSUMPTION LEVELS

	Per family consumption (dollars)	Total consumption (billion dollars)
Regime I	\$12,857.50	\$707.163
Regime II	12,899.90	709.495
Maximum Net Gain	42.40	2.332

Given the hypothetical values of the parameters, the maximum gain from paying interest on cash balances is approximately \$2.3 billion per year. From this gain must be subtracted the real resource costs of instituting a payments mechanism.[5] In the above example, if these costs should exceed \$2.3 billion per year, the society would, of course, be better off if no interest were paid on cash balances.

Given the optimal level of consumption in the two regimes, one can calculate the optimal transactions and optimal inventory stocks in the two regimes. Table 3 presents the figures based on the foregoing assumptions concerning the parameter values.

The effects of paying interest on money become apparent from Table 3. Brokerage transactions are substantially reduced while a larger number of commodity market transactions are undertaken. The supply of cash balances must be increased by over 400 percent with a corresponding decrease in holdings of bonds of about 60 percent. The increase in consumption from regime I to regime II results in part from the fact that nonproductive commodity inventories are reduced and shifted into productive capital. Further gains are achieved through the other components of the net profits from overall inventory management. In order to make the sources of these gains explicit, we present (in Table 4) the total income and expenditure flows under the two regimes.

It is clear from Table 4 that the welfare gain arising from payment of interest on cash balances does not simply come from a higher income on a larger stock of productive capital. The example cited suggests that gains also arise from a reduction in the costs of bond market transactions which outweight both increased commodity market transactions costs and the net increase in overall inventory management costs.

IV. Further Consequences of the Analysis

The suggestion of paying interest on cash balances has led some economists to speculate that individuals would shift their entire financial portfolio out of bonds and into money. Let us, therefore, look at the necessary conditions for bond holdings

[5] In his recent contribution on the optimal quantity of money, Friedman (1969) using a completely different analytic approach, estimated the gross welfare gain from paying interest on money to be between \$2 and \$4 billion per year, assuming a net rate of return on capital of 5 percent. Our own estimates (which are essentially illustrative) suggest a similar gain based on a 6 percent net return on capital. Friedman goes on to estimate the gain using a 17 percent rate of return, and finds the gain to range between \$60–\$100 billion per year. Our own calculations suggest that at a 15 percent rate of return the gain would be approximately \$5–\$6 billion per year.

TABLE 3—OPTIMAL TRANSACTIONS AND INVENTORY HOLDINGS

Regime I: ($r_m = 0$)
Regime II: ($r_m = r_b = r_k - \alpha_k$)

	$(n)^*$ (units)	$(m)^*$ (units)	$(\overline{M})^*$ (dollars)	$(\overline{B})^*$ (dollars)	$(p\overline{Q})^*$ (dollars)	$(p\overline{K})^*$ (dollars)
Per Family						
Regime I	5.28	40.09	952.16	5316.24	160.35	49839.65
Regime II	1.47	59.56	4290.50	2051.16	108.29	49891.71
	(bil. un.)	(bil. un.)	(bil. dol.)	(bil. dol.)	(bil. dol.)	(bil. dol.)
Totals						
Regime I	.290	2.205	52.37	292.39	8.82	2741.20
Regime II	.081	3.276	235.98	112.81	5.96	2744.04

TABLE 4—REVENUES AND EXPENDITURES
(dollars per family)

Source	Regime I		Regime II		Net Change	
Revenues						
Human Income (Y)	$10,000.00		$10,000.00		$ .00	
Income from capital $r_k(p\overline{K})^*$	7,475.95		7,483.76		+ 7.81	
Income from bonds $r_b(\overline{B})^*$	318.97		123.07		−195.90	
Income from money $r_m(\overline{M})^*$	0.00		257.43		+257.43	
Total Revenues		$17,794.92		$17,864.26		+$69.34
Expenditures						
Inventory Management Costs						
$\alpha_k(p\overline{K})^*$	$ 4,485.57		$ 4,490.25		+$ 4.68	
$\alpha_b(\overline{B})^*$	26.58		10.26		− 16.32	
$\alpha_q(p\overline{Q})^*$	9.62		6.50		− 3.12	
$\alpha_m(\overline{M})^*$	9.52		42.91		+ 33.39	
Subtotal	$ 4,531.29		$ 4,549.92		+$18.69	
Transaction Costs						
$\beta_b n$	$ 79.18		$ 21.99		−$57.19	
$\beta_q m$	8.02		11.91		+ 3.89	
Subtotal	$ 87.20		$ 33.90		+$53.30	
Taxes						
Transfer Taxes						
$r_b(\overline{B})^* + r_m(\overline{M})^*$	$ 318.97		$ 380.50		+$61.53	
Cost of Payment Mechanism $h(r_m)$	$ 0.00		$ 0.00		0.00	
Subtotal	$ 318.97		$ 380.50		+$61.53	
Total Expenditures		$ 4,937.46		$ 4,964.32		+$26.26
Net Profit Available for Consumption Expenditures		$12,857.46		$12,899.94		+$42.48

to be positive when interest is paid on money. Given the definition of bond balances,

$$\bar{B} = \frac{pq}{2} - \frac{pq}{2n}, \tag{33}$$

average bond holdings will be positive so long as $n > 1$. When the equilibrium value is substituted for n from equation (29), the condition becomes

$$n^* = \sqrt{\frac{(\alpha_m - \alpha_b)pq}{2\beta_b}} > 1, \tag{34}$$

which can be written as

$$\alpha_m > \alpha_b + \frac{\beta_b}{pq/2} \tag{35}$$

Thus bond holdings will coexist with money holdings even when interest is paid on cash balances—so long as the cost per dollar of money inventories exceeds the cost per dollar of bond inventories plus the cost per dollar of bond transactions.

A similar analysis reveals the conditions required for money balances to be positive in a stationary economy.[6] Since

$$\bar{M} = \frac{pq}{2n} - \frac{pq}{2m}, \tag{36}$$

money balances will be positive so long as $m > n$. Zero money balances will occur when $m = n$, i.e., when there is perfect synchronization between bond market and commodity market transactions. Money balances will be positive so long as

$$\sqrt{\frac{(r_m - \alpha_m + \alpha_q)}{2\beta_q}}\,pq > \sqrt{\frac{(r_b - \alpha_b) - (r_m - \alpha_m)}{2\beta_b}}\,pq, \tag{37}$$

or

$$\frac{(r_m - \alpha_m + \alpha_q)}{(r_b - \alpha_b) - (r_m - \alpha_m)} > \frac{\beta_q}{\beta_b} \tag{38}$$

Equation (38) states that cash balances will be positive so long as the ratio of net profit from holding money rather than commodities, to net profit from holding bonds rather than money, exceeds the ratio of commodity market to bond market transaction costs.

Finally, it has been suggested that if interest payments are made on money, individuals will cease to spend money balances. The foregoing analysis suggests that interest payments on cash balances will induce individuals to hold larger average cash balances, but that they will spend them more frequently by undertaking a larger number of transactions in commodity markets.

V. Summary and Conclusions

The analysis presented above suggests that Pareto-efficiency requires that interest be paid on cash balances which is equal to the rate of return on bonds and the net rate of return on capital. This conclusion, which has also been derived from the traditional utility maximizing model which includes cash balances in the utility function, emerges unscathed when the problem of determining the optimal quantity of money is subjected to an extended inventory theoretic analysis. The inventory theory approach makes clear that the problem of an optimal money supply is inextricably linked to the problem of optimal bond inventories, commodity inventories, and capital stock.

The inventory theory specification of the optimal money problem is extremely rich. It yields explicit demand functions for money, bonds, commodity inventories, and capital, as well as expressions for the optimal number of transactions in different markets.

Moreover, the inventory theory ap-

[6] B. Pesek and T. Saving have claimed that when interest is paid on money, real cash balances go to zero.

proach enables one to examine the explicit sources and magnitudes of the welfare gain resulting from an optimal money policy. The major conclusions concerning the effects of paying interest on money may be summarized as follows:

(a) cash balances will increase,
(b) bond holdings will decrease,
(c) commodity inventory holdings will decrease,
(d) physical reproductive capital will increase,
(e) commodity market transactions will increase,
(f) bond market transactions will decrease,
(g) consumption will increase.

These conclusions are, of course, subject to the qualification that the real resource costs of instituting an interest payment mechanism are lower than the potential gross welfare gain from increased consumption. The analysis clarifies the distinction between a) the virtually costless process of increasing nominal balances and b) the potentially costly process of instituting a payment mechanism on money sufficient to induce individuals to hold the socially optimal collection of asset inventories. One obvious mechanism for paying interest on cash balances would be to induce a steady deflation. Such a policy, however, would be very costly both in terms of transitional adjustments as well as resource costs of continuously changing price tags on commodities. A "second best" solution likely to be less costly in terms of real resources and yet be capable of capturing a substantial portion of the potential welfare gain would be to eliminate the present prohibition of interest payments on demand deposits and to allow explicit payment of interest on bank reserves and vault cash. A third solution would be for banks to be permitted to issue their own competitively produced notes.

The inventory theory approach to the optimal money problem as developed in this paper draws explicit attention to the capital costs of inventory management as well as the capital costs of transacting in bond and commodity markets. It still remains to generalize the model to take explicit account of the human income costs of transactions, by investigating effects of an optimal money policy on the work-leisure margin. Finally the model should be generalized to introduce risk and uncertainty so as to capture the precautionary and speculative motives for holding cash balances.

REFERENCES

W. J. Baumol, "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quart. J. Econ.*, Nov. 1952, *66*, 545–56.

R. Clower, "What Traditional Money Theory Really Wasn't," *Can. J. Econ.*, May 1969, *2*, 299–302.

E. Feige and D. Nichols, "Money, Wealth and Welfare," *Social Systems Research Institute Workshop Series*, Firm and Market No. 6828, Univ. Wisconsin 1968.

M. Friedman, *A Program for Monetary Stability*, New York 1960.

———, *The Optimum Quantity of Money*, Chicago 1969.

H. Johnson, "Inside Money, Outside Money, Income, Wealth and Welfare in Monetary Theory," *J. Mon. Cred. Bank*, Feb. 1969, *1*, 30–45.

D. E. W. Laidler, "Money, Wealth and Time Preference in a Stationary Economy," *Can. J. Econ.*, Nov. 1969, *2*, 526–35.

A. Marty, "Inside Money, Outside Money and the Wealth Effect: A Review Essay," *J. Mon. Cred. Bank.*, Feb. 1969, *1*, 101–11.

D. Patinkin, *Money, Interest and Prices*, New York 1965.

B. Pesek and T. Saving, *Money, Wealth and Economic Theory*, New York 1967.

P. A. Samuelson, "What Classical and Neo-Classical Monetary Theory Really Was," *Can. J. Econ.*, Feb. 1968, *1*, 1–15.

J. Tobin, "The Interest Elasticity of Transactions Demand for Cash," *Rev. Econ. Statist.*, Aug. 1956, *38*, 241–47.

A Utility Theory of Representative Government

By EDWIN T. HAEFELE*

The devolution of economic and political meaning to the individual—hence the identification of the individual *act* as the source of economic and political values—culminated with eighteenth century rationalism. It was then that the foundation was laid for an economic theory which was to dominate the Western world. The political theory of representative government, codified by the American Constitution, was formulated at the same time. Both theories were devised by men who, as Lawrence Frank puts it, ". . . were persuaded by the 18th century belief in the rationality of man and accepted proposals that emphasized the individual's capacity for acting rationally in pursuing his own self-interest and happiness; calculating his prospective gains and losses" (p. 809).

Since that time, the economic theory based on personal utility calculations has prospered while the political theory based on these same calculations has languished and is now suspect. This is not to say that everywhere men still trust a *laissez faire* market economy while mistrusting representative government. It is to say that economic theory, using the concept of a competitive market, can explain much about economic systems both competitive and otherwise, while Anglo-American political theory, having developed no such counterpart concept, has lost its organizing principle.

The loss came very early, for in one lifetime came the brilliant exposition of utility-based political theory, the Federalist Papers; the not-so-"felicific calculus" of Jeremy Bentham; and the fatal utilitarianism of J. S. Mill. Utility-based political theory died aborning, yet the government designed squarely on its precepts has prospered for nearly two hundred years. Prospered because of its utility foundation? Prospered in spite of its utility foundation? Do we know?

In 1835, Alexis de Tocqueville was clear it was because of its utility foundation. "If . . . you do not succeed in connecting the notions of right with that of personal interest, which is the only immutable point in the human heart, what means will you have of governing the world except by fear?" (pp. 147–48). The authors of the Federalist Papers were likewise convinced. Lately the question has shifted considerably and split into two parts: does representative government really *have* a utility base, and is a utility base appropriate for the problems of the present? Both questions have important practical as well as theoretical significance and it may prove useful to review why that is so.

Aggregate measures, whether they be *GNP*, personal income, net social return on investment, or whatever, are increasingly disputed because of the distribution or incidence problem. Who benefits and who pays now occupy the attention of

* Resources for the Future, Inc. Special thanks are due to Allen V. Kneese and Elizabeth Duenckel of the RFF staff for advice, counsel, and much hard work. A note of appreciation is due to Eleanor B. Steinberg, who convinced me not to publish an earlier draft. Orris Herfindahl, Dennis Mueller, and Robert S. Steinberg read the manuscript and made helpful suggestions for improvement. I am also indebted to one of the reviewers for the Farquharson and Shapley references.

economists as they turn to questions of public goods. Welfare economics has unearthed no answer to the question of inter-personal comparisons of utility in the absence of a social welfare function. The generation of such functions has been delegated to the political process, since it is the formal social choice mechanism. Yet economists have found few political scientists interested in undertaking the task of explaining precisely how the generation occurs. Lacking the explanation, economists have tried their hand at it themselves, notably in the work of James Buchanan and Gordon Tullock (1965) and Anthony Downs.

Another economist, Kenneth Arrow (1951, 1963) had already explained how it is *not* done and that explanation has had more impact than the attempts to explain how it *is* done.[1] In brief, Arrow was interested, as were the Founding Fathers, in determining collective or social choices (choices of public policies or choices among candidates for office) on the basis of individual (voter) preferences. Arrow set up four seemingly reasonable conditions which a social choice mechanism should be expected to meet and found that, as a general proposition, no such mechanism could be devised. No one has disproved Arrow's Possibility Theorem, although some, particularly Duncan Black (Oct. 1969), have recently raised some hard questions about the relevance of Arrow's conditions.

Arrow had two parts to his Possibility Theorem and much more ink has been spilt about the General Possibility Theorem than about the Possibility Theorem for Two Alternatives. Arrow proved that the method of majority rule applied to two alternatives does satisfy the four conditions and that the Possibility Theorem for Two Alternatives was, "in a sense, the logical foundation of the Anglo-American two-party system" (1951, p. 48). The restricted Theorem has been largely ignored by later writers, perhaps because it was considered trivial from a social welfare function viewpoint.

Arrow did not pursue the implications of his restricted Theorem. His concern and the concern of most economists have been for the more general problem and, implicit in such concerns, the analogy with general equilibrium theory in economics. Such concerns and analogy are almost demanded by the logic of individual preference *orderings* and a social preference *ordering* by themselves.

Such orderings are incomplete postulates in political terms. For example, take the well-known voting paradox ordering

A	*B*	*C*
B	*C*	*A*
C	*A*	*B*

by which, with majority vote as the decision rule, no social ordering can be found. Implicit in that judgment is some voting matrix to which the preference orderings can be related. One such voting matrix is as follows (Y = yes vote, N = no vote, subscripts indicate ordinal ranking of ABC by each voter):

Issue	Voter I	Voter II	Voter III
A	Y_1	N_3	Y_2
B	Y_2	Y_1	N_3
C	N_3	Y_2	Y_1

Here we assume each voter will vote for either his first or second choice, but never for his third choice. This assumption preserves the cyclical outcome for pairwise choices, *A* preferred to *B*, *B* preferred to *C*, and *C* preferred to *A*.

Suppose, however, another voting matrix which does no violence to the ordinal ranking of ABC by each voter:

[1] I am ignoring here the distinction made elegantly by Paul Samuelson between a Bergson social welfare function and Arrow's constitutional function.

Issue	Voter		
	I	II	III
A	Y_1	N_3	N_2
B	Y_2	Y_1	N_3
C	Y_3	N_2	Y_1

Now it seems clear that B will win, since Voter I is willing to vote for A, B, or C, and in so doing he provides the winning margin for either B or C. He prefers B to C.

Yet another voting matrix with the same ordinal[2] ranking:

Issue	Voter		
	I	II	III
A	N_1	Y_3	Y_2
B	Y_2	N_1	Y_3
C	Y_3	N_2	Y_1

Now it appears that Voters I and II will trade votes, defeating both A and B and allowing C to win. This results because both Voters I and II have, as a first choice, the *defeat* of an alternative. Thus, though ABC are mutually exclusive alternatives, trading is possible given the appropriate voting matrix.

These cases are commonplace occurrences in the political arena but they can not be examined by looking at preference orderings alone. In particular, the analogy of a market is particularly inapposite in the last example. There is no market mechanism by which I can directly express the intensity of my dislike for a product. The social choice mechanism of representative government combines a vote matrix with preference orderings to allow expression of negative intensities as well as positive intensities of preference. The use of the combined matrix will be shown later to be of value in understanding the utility mechanism.

Another reason for economists' growing interest[3] with collective or social choices is the increasing problem of externalities in the production process, particularly when these are harmful to the public. Air and water pollution resulting from man's economic activities are the most visible example. If air and water are free goods, or nearly so, then they may be overused because private cost calculations do not include the costs imposed on other people. Hence, as Buchanan and Tullock have noted,[4] the economic basis for taking collective actions has shifted from opportunities for external economies to conflicts over external diseconomies. The import of this shift is to focus attention on governmental action rather than on corporate action. For, while schemes can be devised, effluent charges for example, which would internalize these costs, they must be adopted to be effective. Adoption presumes a collective choice and persons hurt by this choice will resist it. If the scheme is rejected, how are we to judge the rejection? Is whatever the political process churns out "right" by definition? Some social scientists have called this trust of the political system into question. Daniel Bell asks for a new political theory to provide a way of choosing between the welfare gains and the welfare losses.[5]

[2] Note, however, that now the ranking must be strictly interpreted as intensity of preferences, whether for or against. Thus Voter I prefers first, A to lose; second, B to win; and third, C to win.

[3] See particularly the work of Robert Ayres and Allen Kneese and that of Lloyd Shapley and Martin Shubik.

[4] ". . . as people get richer, they need to rely less and less on their neighbors to cooperate in securing the individual benefits of possible joint activities, but they may need to rely more and more on some collective mechanism to prevent themselves, and their neighbors, from imposing mutually undesirable costs on each other. . . . 'congestion' replaces 'cooperation' as the underlying motive force behind collective action" (1965, p. 69).

[5] "The political tradition from John Locke to Adam Smith paved the way for a new society in which representative government and the free market economy served as the framework for a system of individual decision-making based on self-interest and rational choice. Can one write a new political theory . . . that deals with a service state and a society characterized by a new mixture of individual and communal public and private

Allen Schick warns that the political system may be defective in a manner analogous to an imperfect market. He notes that since the impact of public goods decisions falls unequally on different groups, the political mechanism, far from providing clear welfare criteria for choice, may produce either too much of a public good (defense?) or too little (environmental quality?). Is there no political analogue to Smith's invisible hand?

In sum, determining whether or not representative government does have a utility base and asking whether or not such a base is appropriate are both revelant areas for study. This paper is mainly concerned with demonstrating the utility base but does offer some thoughts on the latter question.

I. Representation and Individual Utility

The link between utility and representation has proved a major stumbling block which hinders our understanding of the utility base of representative government. That representative democracy, as opposed to pure democracy, was necessary for effective government was almost self-evident in the 18th century.

James Madison in Federalist Paper No. 55 expresses the point most succinctly, "Had every Athenian citizen been a Socrates, every Athenian assembly would still have been a mob." Need it be added that Madison's words relate to information costs, revealed preferences, and the lack of a vote-trading mechanism, or that modern proposals that everyone vote on all issues by electronic processes suffer the same defect as the Athenian assembly?

Arguing the necessity of representation on negative grounds does not address the question of the link between individual utility and the representative, however. Furthermore, the contention that the representative will be wiser, more judicious, less swayed by whims of the moment—whatever its truth may be—likewise begs the question. Let us address it directly.

Consider three men as comprising a district. Two independent issues are posited as being important for resolution, and the men's positions on these issues are described by the following combined vote and preference matrix:[6]

Issue	Voter I	Voter II	Voter III
A	Y_2	Y_2	N_2
B	N_1	Y_1	Y_1

where again Y is a vote for, N a vote against, and the subscripts are the ordinal rankings of the issues by each man. In this case, all three men rank issue B as more important than issue A, Voter I prefers the defeat of issue B to passage of issue A, Voter II prefers the passage of B to the passage of A, and Voter III prefers the passage of B to the defeat of A.

Two methods exist whereby these men may decide these issues. They may meet as an assembly; in such case, it is obvious that under majority rule, both issues will pass (no trades are possible). Alternatively, they may elect a representative (not one of the three voters). In that case they face a mutually exclusive choice of one of four possible outcomes on the two issues. If we display these as alternative outcomes with the consequences for each voter under each outcome we have: (P = pass, F = fail)

decision-making units? . . . where joint decisions are to be made, are there clear welfare criteria that justify one choice rather than another?" (pp. 699, 977).

[6] In this and all subsequent examples, I assume no position on whether a representative should lead or follow his constituency. The individual preference orderings and voting positions can be considered sui generis or as having been formed from the persuasions of prospective or actual representatives.

Issue *A*	*P*	*P*	*F*	*F*
Issue *B*	*P*	*F*	*P*	*F*
Voter I wins	2	12	none	1
Voter II wins	12	2	1	none
Voter III wins	1	none	12	2

Thus, if a representative were elected on a $\left[\begin{smallmatrix}P\\P\end{smallmatrix}\right]$ platform, Voter I would win only his second choice (passage of *A*), Voter II would win both his first and second choices, and Voter III, his first choice only. Obviously, Voter II would vote for a representative espousing a $\left[\begin{smallmatrix}P\\P\end{smallmatrix}\right]$ platform, but Voters I and III would be motivated to look for alternatives to $\left[\begin{smallmatrix}P\\P\end{smallmatrix}\right]$. No single alternative is preferred by both voters. We may conclude that $\left[\begin{smallmatrix}P\\P\end{smallmatrix}\right]$ would win over any rival, and that the perceptive aspirant for office will run on this platform.

Sometimes an assembly outcome depends on vote-trading in which issues of lesser utility are traded for issues of greater utility. Consider this rearrangement of preferences:

	Voter		
	I	II	III
A	Y_2	Y_1	N_1
B	N_1	Y_2	Y_2

The assembly outcome of this arrangement would be that issue *A* passes and issue *B* fails. This results from an attempted trade between Voters I and III (which would result in failing both issues), which Voter II prevents by giving up issue *B* (voting *N* instead of *Y*), thereby keeping Voter I's *Y* vote on issue *A*.

Under an election process, obviously the voters cannot directly trade votes. They are again faced with a choice of mutually exclusive alternatives. The alternative outcomes are:

Issue *A*	*P*	*P*	*F*	*F*
Issue *B*	*P*	*F*	*P*	*F*
Voter I wins	2	12	none	1
Voter II wins	12	1	2	none
Voter III wins	2	none	12	1

Now, if we are to arrive at the same outcome $\left[\begin{smallmatrix}P\\F\end{smallmatrix}\right]$ as did the men when meeting as an assembly, the path which the choice process takes becomes crucial. Proceeding as we did before, Voters I and III can be attracted to $\left[\begin{smallmatrix}F\\F\end{smallmatrix}\right]$ as an alternative to $\left[\begin{smallmatrix}P\\P\end{smallmatrix}\right]$. Were $\left[\begin{smallmatrix}F\\F\end{smallmatrix}\right]$ to be chosen by one candidate, a second candidate could win on a $\left[\begin{smallmatrix}P\\F\end{smallmatrix}\right]$ platform, *assuming no other candidate runs.*

While it is not remarkable that a path can be found, by means of a chain of assumptions, which brings us to the solution reached through assembly trading, it is significant that the chosen path is not unlike a two-party system groping toward positions on issues.

Before formulating rules by which that path can be specified when more than two issues are involved, it may prove helpful to explore outcomes of all ordinal permutations of the $\left[\begin{smallmatrix}YYN\\NYY\end{smallmatrix}\right]$ voting matrix. These are shown in Table 1.

If we view each permutation as a separate district, we can see what happens when voters in different districts are not concerned about the *same* issues.

Suppose that the voters from districts (permutations) 4, 5, and 6 are now considered in terms of three issues, as follows:

	Issue	Voter		
		I	II	III
District 4	*A*	Y_1	Y_2	N_2
	B	N_2	Y_1	Y_1
	C	3	3	3
District 5	*A*	Y_2	Y_1	N_1
	B	3	3	3
	C	N_1	Y_2	Y_2
District 6	*A*	3	3	3
	B	Y_2	Y_2	N_1
	C	N_1	Y_1	Y_2

If an issue is not relevant in a district—as, for example, issue *C* in district 4—we can ascribe to it the third position in the ordinal ranking of each voter even though we ascribe no *Y* or *N* position. Since we

TABLE 1—LEGISLATIVE-ASSEMBLY OUTCOMES
2x3 Matrix

*Case 12	Voter			Outcomes	
Issue	I	II	III	Assembly	Representative
A	*Y*	*Y*	*N*		
B	*N*	*Y*	*Y*		
	Ordinal permutations:				
(1)	1	1	1	Pass	Pass
	2	2	2	Pass	Pass
(2)	1	2	1	Pass	Pass
	2	1	2	Pass	Pass
(3)	1	1	2	Pass	Pass
	2	2	1	Pass	Pass
(4)	1	2	2	Pass	Pass
	2	1	1	Pass	Pass
(5)	2	1	1	Pass	Pass
	1	2	2	Fail	Fail
(6)	2	2	1	Fail	Fail
	1	1	2	l ass	Pass
(7)	2	1	2	Pass	Pass
	1	2	1	Pass	Pass
(8)	2	2	2	Pass	Pass
	1	1	1	Pass	Pass

Note: *Case designations are formed by counting the frequency with which voters appear in the initial coalitions. In Case 12, the only nontrivial 2x3 case, one voter is in two coalitions, and two voters are in one coalition, hence Case 12.

have not changed the voting pattern or the ordinal rankings, we know the outcome in each district (whether decided by election of a representative or by assembly action) by reference back to Table 1. Thus the representatives from each district, when they meet in a regional assembly, could be shown as

Issue	Representative		
	4	5	6
A	Y_2	Y_1	3
B	Y_1	3	N_2
C	3	N_2	Y_1

If we use majority rule as the decision rule, then the blanks may be replaced, plausibly, by *N*'s, since the outcome is not changed thereby. The resulting matrix

Issue	Representative		
	4	5	6
A	Y_2	Y_1	N_3
B	Y_1	(N_3)	N_3
C	N_3	N_3	Y_1

has one potential trade (which is blocked by Representative 5) and the outcome is to pass the first two issues and fail the third.

Were the nine voters to come together as an assembly, thus

Issue	Voter								
	I	II	III	IV	V	VI	VII	VIII	IX
A	Y_1	Y_2	N_2	Y_2	Y_1	N_1	3	3	3
B	N_2	Y_1	Y_1	3	3	3	Y_2	Y_2	N_1
C	3	3	3	N_1	Y_2	Y_2	N_1	Y_1	Y_2

and the blanks replaced by *N*'s

Issue	Voter								
	I	II	III	IV	V	VI	VII	VIII	IX
A	Y_1	Y_2	N_2	Y_2	Y_1	N_1	N_3	N_3	N_3
B	N_2	Y_1	Y_1	N_3	N_3	N_3	Y_2	Y_2	N_1
C	N_3	N_3	N_3	N_1	Y_2	Y_2	N_1	Y_1	Y_2

again the first two issues pass and the third fails, although the process of trading which accomplishes this outcome is a little more complex. All issues are losing initially, but Voter VII must change his vote on issue *A* to block a trade between Voters I and III which would cause issue *C* to win. Similarly, Voter IV must change his vote on issue *B* to keep issue *C* from winning because of a possible trade between Voter III and either Voter V or VI.

While the foregoing illustration may be obvious, it does demonstrate that the process works across districts without uniformity of issues in every district.

Rules for Solution in the 3x3 Vote Matrix

Following William Riker in using only minimum winning coalitions, only two 3x3 vote patterns need be considered:

Case 030				Case 111		
Y	*Y*	*N*		*Y*	*Y*	*N*
Y	*N*	*Y*		*Y*	*Y*	*N*
N	*Y*	*Y*		*Y*	*N*	*Y*

Each case has 216 ordinal permutations. (Manipulation of *Y*'s and *N*'s are not necessary to exhaust the set of permutations.)

Legislative Vote Trading

Rules for solving the cases considered as assemblies are fairly straightforward. We assume voting stances are known but that relative importance of issues (ordinal ranking) is revealed only by actions. The rules are:

1) Trades take place if, and only if, they are mutually advantageous.

2) Any trader prefers a higher gain to loss ratio to a lower one, e.g., will trade a third choice for a first choice in preferences to a second choice for a first.

3) Any trade can be reversed (cancelled) by a third voter if he can offer one of the traders an alternative which is more advantageous to that trader and less damaging to himself than the trade would be.

4) Bluffs and threats, defined as actions which, if taken, would harm the actor, are not allowed.

5) While only ordinal utilities are used, it is convenient to set choice 1 = choices 2+3 for all voters.

6) Majority rule is the decision rule.

Solutions determined by these trading rules were assumed "correct" for the purpose of devising rules for selection of a representative.

Rules for Selecting a Representative

Again we assume that voting stances (opinion polls?) are known at the outset. The rules determine the decision path:

1) Start at the nominal outcome, i.e., count the votes. As presented here, that is always at the vector $\begin{bmatrix} P \\ P \\ P \end{bmatrix}$

2) Select from the remaining seven possible vectors, the outcome(s) which is (are) most advantageous:

(a) to the two voters "worst off" in the $\begin{bmatrix} P \\ P \\ P \end{bmatrix}$ vector[7]

(b) or, if the two "high" men are tied, to the one low man and either one of the high men[8]

[7] Example: Given a vote and ordinal matrix

Y_2	Y_1	N_1
Y_3	N_2	Y_2
N_1	Y_3	Y_3

hence possible outcomes

	P P P	P F P	F P P	F F P	P P F	P F F	F P F	F F F
Voter I wins	23	2	3	—	123	12	13	1
Voter II wins	13	123	3	23	1	12	—	2
Voter III wins	23	3	123	13	2	—	12	1

Voters I and III are "worst off." The only vector which is mutually advantageous is $\begin{bmatrix} F \\ P \\ F \end{bmatrix}$ (If more than one vector is possible, all are chosen.) Keep in mind that the ordinal ranking measures the importance of the issue to the voter. Thus $\begin{bmatrix} Y_2 \\ Y_3 \\ N_1 \end{bmatrix}$ is interpreted that the voter prefers first the defeat of *C*, second the passage of *A*, and third the passage of *B*. If the outcome is $\begin{bmatrix} P \\ P \\ P \end{bmatrix}$ this voter wins his second and third choices and not his first.

[8] Example:

P P P	P F P	F P P	F F P	P P F	P F F	F P F	F F F
13	3	1	—	123	23	12	2
13	123	3	23	1	12	—	2
23	3	123	13	2	—	12	1

The vector selected would be $\begin{bmatrix} F \\ P \\ F \end{bmatrix}$

(i) if more than one vector choice is possible, choose the one most favorable to the low man[9]

(ii) if no such vector exists, choose the outcome favored by the two high men[10]

(c) or, if all three voters are tied, to any pair of voters.[11]

3) If no vector can be chosen under (2), then the initial vector is chosen by all parties.

4) If a vector is selected under (2), then it becomes one party position.

5) One other party position is selected from the remaining six vectors by choosing the vector(s) which adds, and only adds, to the winnings of one member of the first party (in 2) above). If more than one outcome is possible, choose the outcome most advantageous to the voter not a member of the first party.[12]

6) The two party positions are matched and the vector selected which would receive the majority vote. (As set up here, this will always be the vector selected in 5) above.)

7) While only ordinal utilities are used, it is convenient to set choice 1 = choices 2+3 throughout for all voters.

8) Majority vote is the decision rule.

Results of applying these rules to the 3x3 matrix are shown in Table 2.[13] Several items in the results need some explanation. First, where a bargaining situation develops (either in the legislative process or in the selection process), the same assumption must be used if the two processes are to produce the same result. This is not surprising, but does mean that it is occasionally possible, particularly in Case 111, for some of the outcomes to be interpreted differently. Second, there are two permutations of Case 030 for which two solutions are possible because of our "choice 1 = choices 2+3" rule.

These two cases are the last vestige of the cyclical majority phenomenon. They are testimony to the sturdy truth underlying Arrow's General Therorem.

[9] Example:

P	*P*	*F*	*F*	*P*	*P*	*F*	*F*
P	*F*	*P*	*F*	*P*	*F*	*P*	*F*
P	*P*	*P*	*P*	*F*	*F*	*F*	*F*
23	2	3	—	123	12	13	1
13	123	3	23	1	12	--	2
13	3	123	23	1	--	12	2

The vector selected would be $\begin{bmatrix} P \\ F \\ F \end{bmatrix}$

[10] Example:

P	*P*	*F*	*F*	*P*	*P*	*F*	*F*
P	*F*	*P*	*F*	*P*	*F*	*P*	*F*
P	*P*	*P*	*P*	*F*	*F*	*F*	*F*
13	1	3	—	123	12	23	2
13	123	3	23	1	12	—	2
23	3	123	13	2	—	12	1

The vector selected would be $\begin{bmatrix} P \\ F \\ F \end{bmatrix}$

[11] Example:

P	*P*	*F*	*F*	*P*	*P*	*F*	*F*
P	*F*	*P*	*F*	*P*	*F*	*P*	*F*
P	*P*	*P*	*P*	*F*	*F*	*F*	*F*
23	2	3	—	123	12	13	1
23	123	3	13	2	12	—	1
23	3	123	13	2	—	12	1

The vector selected would be $\begin{bmatrix} P \\ F \\ F \end{bmatrix}$

[12] Example:

P	*P*	*F*	*F*	*P*	*P*	*F*	*F*
P	*F*	*P*	*F*	*P*	*F*	*P*	*F*
P	*P*	*P*	*P*	*F*	*F*	*F*	*F*
23	2	3	—	123	12	13	1
13	123	3	23	1	12	—	2
23	3	123	13	2	—	12	1

Vector $\begin{bmatrix} F \\ P \\ F \end{bmatrix}$ is chosen by one party. Two vectors dominate it under the add and only add rule: $\begin{bmatrix} P \\ P \\ F \end{bmatrix}$ and $\begin{bmatrix} F \\ P \\ P \end{bmatrix}$

The vector selected is $\begin{bmatrix} P \\ P \\ F \end{bmatrix}$

[13] All cases were solved by hand, but to avoid the possibility of a shifting premise, computer programs were written and the cases solved again. Fortunately, the two methods give identical answers. Elizabeth Duenckel developed the programs and her help is gratefully acknowledged.

TABLE 2—LEGISLATIVE-REPRESENTATIVE OUTCOMES
3x3 Matrix

*Case 030	Voter			Outcomes		*Case 111	Voter			Outcomes	
Issue	I	II	III	Legislative	Representative	Issue	I	II	III	Legislative	Representative
A	*Y*	*Y*	*N*			*A*	*Y*	*Y*	*N*		
B	*Y*	*N*	*Y*			*B*	*Y*	*Y*	*N*		
C	*N*	*Y*	*Y*			*C*	*Y*	*N*	*Y*		
Ordinal permutations (216 possibilities):						Ordinal permutations (216 possibilities):					
100 of which				Pass Pass Pass	Pass Pass Pass	132 of which				Pass Pass Pass	Pass Pass Pass
38 of which				Pass Pass Fail	Pass Pass Fail	42 of which				Pass Pass Fail	Pass Pass Fail
38 of which				Pass Fail Fail	Pass Fail Fail	21 of which				Pass Fail Pass	Pass Fail Pass
38 of which				Fail Pass Pass	Fail Pass Pass	21 of which				Fail Pass Pass	Fail Pass Pass
2 of which				{Fail, Fail, Fail} or {Pass, Pass, Pass}	{Pass, Pass, Pass} or {Fail, Fail, Fail}						

Note: *Cases are designated as in Table 1. Cases 030 and 111 are the only two nontrivial 3x3 cases.

II. The Cyclical Majority Uncycled[14]

While the paradox of the cyclical majority has been traditionally posed as involving mutually exclusive issues—alternative social states, candidates for one post, considered without side payments—both side payments and independent issues have been introduced into recent discussion. Arrow comments on side payments:

> If, instead of assuming that each individual votes according to his preferences, it is assumed that they bargain freely before voting (vote-selling), the paradox appears in another form. . . . If a majority could do what it wanted, then it would be optimal to win with a bare majority and take everything; bu any such bargain can always be broke up by another proposed majority. [196(p. 61]

We have seen, however, that o vote-trading is allowed, the paradox d not hold if issues are independent. T result is not limited to single-pea preference functions, which Tullock : Paul Simpson have investigated; nor n it come about because of consideration utility under uncertainty, as in the w of Coleman; nor does it depend on existence of a pseudo-price mechanism posited by Robert Wilson.

To demonstrate this, let us examine permutation usually considered in a ı tually exclusive issue context as an in pendent issue case (using our rules

[14] James C. Coleman read an earlier version of this section and made valuable suggestions for improvement.

legislative trading). Changing the vote matrix to comply with the assumption that each voter votes for his first choice and against the other two, we will have (majority rule):

Issue	Voter I	Voter II	Voter III	Outcome
A	Y_1	N_2	N_3	F
B	N_2	N_3	Y_1	F
C	N_3	Y_1	N_2	F

In order to examine the trading sequence, it will be useful to display the subscripts (rankings) separately and construct a trading matrix:

Issue	Voter I	Voter II	Voter III
A	1	−2	−3
B	−2	−3	1
C	−3	1	−2

Each column vector shows a voter's trading desires. Applying our rules for trading, the voter *can* trade any cell with a negative sign but he will, obviously, only trade lower preferences for higher. Thus in this display each preference vector shows a desire and ability to trade either a second or third choice for a first choice. Choosing a first trade arbitrarily since none is dominant, the trading sequence is illustrated in Table 3.

Since there is no familiar notation to show the trading sequence, it may be helpful to describe the process shown in Table 3. Starting in Tableau I, and recalling that columns represent voters and rows are issues, we see that all three issues fail initially. Voters II and III trade votes on issues B and C (shown by the crossed arrows). This trade, if allowed to hold, results in the solution S_1, namely that, Voter I would win nothing, Voter II would win on his first and second preferences, and Voter III would win on his first and third preferences.

This trade is not stable, however, as it is in Voter I's best interest (and in his power) to reverse the trade by giving up his vote on issue C to Voter II so that Voter II does not need to trade with Voter III. Hence, the S_2 solution at the bottom of Tableau I.

Tableau II is generated by the S_2 solution, i.e., the issues are now F, F, P, and Voter I is shown voting Y on issue C. The negative and positive signs in the trading matrix are changed accordingly and additional trades are sought. There is a possible trade (again shown by crossed arrows) and this time the potential reversal cannot occur. Voter II could reverse this trade only if he is willing to give up Voter I's support on issue C. But, since issue C is his first choice, he will not be willing. Hence, S_3 is shown as the result of the trade between Voters I and II on issues A and B.

Tableau III is generated from this solution (S_3); all issues are now passing, and the new trading matrix shows no further trades are possible. Three votes are shown "circled," indicating that they are cast Y as the result of trades. The result does not depend on Voter I being willing to trade off *both* issues B and C to get A. In other words, $A \not> B+C$. He trades off C to keep B. When he has B, he can use it to get A, which he prefers to B.

When the solution $\begin{bmatrix} P \\ P \\ P \end{bmatrix}$ has been reached, it is stable unless the rule "choice 1 = choice 2+3" is relaxed. If at least two voters think choices $2+3>1$, the solution could be forced to $\begin{bmatrix} F \\ F \\ F \end{bmatrix}$

When this permutation appears in a mutually exclusive context (representative rules) the same two solutions appear. Displaying the outcomes as before:

P	*P*	*F*	*F*	*P*	*P*	*F*	*F*
P	*F*	*P*	*F*	*P*	*F*	*P*	*F*
P	*P*	*P*	*P*	*F*	*F*	*F*	*F*
1	12	—	2	13	123	3	23
1	13	12	123	—	3	2	23
1	—	13	3	12	2	123	23

and noting that $\begin{bmatrix} F \\ F \\ F \end{bmatrix}$ is the initial vector, we can choose no other vector under rule 2. Hence (under rule 3) $\begin{bmatrix} F \\ F \\ F \end{bmatrix}$ is chosen by both parties. Only if choice 1 > choices 2+3 could $\begin{bmatrix} P \\ P \\ P \end{bmatrix}$ dominate $\begin{bmatrix} F \\ F \\ F \end{bmatrix}$.

Thus the legislative rules choose $\begin{bmatrix} P \\ P \\ P \end{bmatrix}$ and the representative rules choose $\begin{bmatrix} F \\ F \\ F \end{bmatrix}$. This divergence is uniquely determined by the convenience rule, choice 1 = choices 2+3. If this rule is relaxed, the two processes

TABLE 3—TRADING SEQUENCE

	Vote Matrix				Trading Matrix			Winning Preferences				
	Voter I	Voter II	Voter III	Outcome	Voter I	Voter II	Voter III			Voter I	Voter II	Voter III
					Tableau I							
Issue A	*Y*	*N*	*N*	*F*	1	−2	−3	Initial Solution	S_0	2nd 3rd	2nd 3rd	2nd 3rd
B	*N*	*N*	*Y*	*F*	−2	−3 (traded with C III)	1 (traded with C II)	(trade)	S_1	none	1st 2nd	1st 3rd
C	*N*	*Y*	*N*	*F*	−3 →	1	−2	(reverse)	S_2	2nd	1st 2nd 3rd	3rd
					Tableau II							
Issue A	*Y*	*N*	*N*	*F*	1 (traded with B III)	−2	3 (traded with B I)					
B	*N*	*N*	*Y*	*F*	−2	−3 ↛[a]	1	(trade)	S_3	1st	1st	1st
C	(*Y*)	*Y*	*N*	*P*	3	−1	2					
					Tableau III							
Issue A	*Y*	*N*	(*Y*)	*P*	−1	2	3					
B	(*Y*)	*N*	*Y*	*P*	2	3	−1	Final	S_F	1st	1st	1st
C	(*Y*)[b]	*Y*	*N*	*P*	3	−1	2					

Note: ╳ indicates trade. → indicates a reversal of trade. ↛ indicates a blocked reversal.

Each successive tableau of voting matrix and trading matrix is set up on the basis of trading done in preceding tableau; thus Tableau II starts at Solution S_2 and Tableau III at S_3.

[a] Reversal blocked because Voter I can retaliate by withdrawing his support on Issue C (last row).

[b] This vote remains a *Y* because it enables Voter I to hold Issue B (second row) so that he can trade it to Voter III.

chose the same outcome. If choice 1 $>$ choices 2+3, $\begin{bmatrix} P \\ P \\ P \end{bmatrix}$ is chosen. If choice 1 $<$ choices 2+3, $\begin{bmatrix} F \\ F \\ F \end{bmatrix}$ is chosen. This convergence requires an addition to the representative rules to deal with the following special case when it occurs.

$$\begin{bmatrix} P \\ P \\ P \\ \hline 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} F \\ F \\ F \\ \hline 2\ 3 \\ 2\ 3 \\ 2\ 3 \end{bmatrix}$$

A more general solution to this problem would require additional specified elements in each ordinal ranking. That was not attempted.

The other nominally cyclical examples (there are a total of 12 in each 216 permutations) come to a single stable solution both as independent issues (legislative rules) and as mutually exclusive choices (representative rules) and are considerably less complicated than the example presented here.

A caveat from the real world should be issued at this point. The practice of statecraft cannot always depend on legislative vote-trading or a two-party choice process. When decisions must be made rather separately, so that vote-trading possibilities are reduced, the decision is rarely made by simple majority vote. The role of the nominating committee as a screening device is familiar to all.[15] It was not by accident that nominating committees, party caucuses, the King's Council, and other devices grew up in Anglo-Saxon government to prevent indecision (lack of convergence) in governmental processes. This growth took place, not in the grip of utility theory as did the "democratic" elements of the system, but as a historical reaction developed during the long and bloody struggle for control of the English Crown. The early history of the English Commons shows it assenting to or rejecting a proposed levy. As the initiative gradually passed from King and Council to Commons, the utility part of the decision process was confined to adjustments on local issues. Individual members' preferences do not shape the bills involving national (rather than local) interests. The latter bills were (and are) shaped by party leadership.

W. Bagehot, writing near the middle of the last century, is still instructive on the point:

> . . . the principle of Parliament is obedience to leaders. . . . The penalty of not doing so is the penalty of impotence. It is not that you will not be able to do any good, but *that you will not be able to do anything at all.* If everybody does what he thinks right, there will be 657 amendments to every motion, and none of them will be carried or the motion either. [p. 141]

Moreover, there are times when even party leadership will not suffice. Writers from Aristotle—"the nature of a *polis* is to be a plurality"—to Coleman[16] have recognized that single issues considered *in vacuo* cannot be resolved by political means.

The knowledge that simple majority rule or individual utility concepts cannot be used in *all* matters of statecraft is no

[15] Lewis Carroll's struggles with majority rule in the election of new officers at Christ Church College, Oxford, are recounted by Black (1963). C. P. Snow in *The Masters* gives an accurate picture of the undemocratic reality.

[16] ". . . there is evidence to suggest that when a single decision dominates a political or social system, . . . the decision process breaks down; and not only is there no "social welfare function," there is overt conflict . . . " (p. 1116).

different, in principle, than the knowledge that the market cannot be trusted to establish a competitive price in a monopoly situation.

III. Political Parties and Personal Utility

Earlier, Arrow's statement that the logical foundation of the Anglo-American two-party system could be found in his Theorem for Two Alternatives was mentioned in connection with the neglect with which the Restricted Theorem has been treated. It should now be clearer why the theorem is not trivial. If the two alternatives are not randomly chosen, but rather are those two positions, which, when put to a vote will result in the same choice as that chosen by the voters if they engage in vote-trading; *then a method of passing from individual tastes to social preferences, excluding inter-personal comparisons of utility, defined for a wide range of sets of individual orderings, neither dictatorial nor imposed, is representative government with a "satisfactory" two-party system.*

The word satisfactory in the above statement refers to whether or not the party system does tend to choose the two positions referred to above. Judging that is no simple task. Just as the price of gasoline at the corner station is no test of the efficacy of our market economy, neither is one issue, one time period, or one area a test of whether a two-party system is correctly reflecting utilities.

The proposition, and the rules on which it is based, have some strong implications. First, they suppose two parties and only two. For whatever reasons adopted, the single member, single vote constituencies in the United States exert a powerful force in that direction in any district or state, while the state electoral system of voting for president is a strong force for the two parties to be nationwide.[17] Second, they suppose non-doctrinaire parties, capable of changing positions to win voter approval. While often deplored (but not by politicians), the American party system qualifies on that count. Third, they suppose only a limited ability of parties to change course once committed in a particular election and less than total information of people's interests. These qualities are approximately present in nature. Fourth, they suppose majority rule to have utility for individuals. A recent discussion and proof of this proposition is given by Douglas Rae and Michael Taylor. Fifth, and finally, they suppose that the results from the permutations in the 3x3 cases are indicative of more general results. No proof is established of this supposition, although it should not be assumed that the set of issues which is important for voter preferences is much larger than three. Regression equations which explain voter and legislator behavior typically contain no more than four or five significant independent variables.[18]

The proposition and rules for electing a representative are in sharp contrast to any form of proportional representation, a system for reflecting every sizable shade of opinion in the legislature.[19] Yet the rules produce the same outcome as would occur if *everyone* were in the legislature.

IV. Optimality Considerations

If trading on independent issues and selection of candidates by their stand on issues are to proceed along optimal lines, then the issues must be "correctly" specified. Any issue, say, federal aid to education, may be framed in hundreds of different ways. How it is framed determines

[17] Contrary to popular and Congressional opinion, the proposed change from awarding electoral votes on a winner-take-all basis, state-by-state to a simple nationwide vote count may greatly imperil the two-party system at the national level.

[18] See the work of Gerald Kramer and of John Jackson.

[19] See Duncan Black (1969) and the work of Ruth Silva for modern treatments of this perennial issue.

not only which people are for it and which against (specifying the vote matrix), but also the intensities of feeling pro and con (thus providing an input to the ordinal matrix). Control over how issues are to be framed is a powerful lever; one that is almost analogous to controlling the initial distribution of income in a market equilibrium analysis.

Let us illustrate this problem by first examining an ordinal matrix in which a legislature of five members is considering Issue A. Each member has his own version (bill) on this issue, as follows:

Issue	Members 1	2	3	4	5
A_1	1	2	3	5	4
A_2	2	1	2	4	5
A_3	3	3	1	3	3
A_4	5	4	5	1	2
A_5	4	5	4	2	1

It is not possible to solve this matrix without implicitly assuming some vote matrix. Before specifying the vote matrix, however, note the possibility of partitioning this matrix on affinity lines (which could correspond to party lines, liberal-conservative lines, urban-rural lines):

Issue	Members 1	2	3	4	5
A_1	1	2	3	5	4
A_2	2	1	2	4	5
A_3	3	3	1	3	3
A_4	5	4	5	1	2
A_5	4	5	4	2	1

Members 1, 2, and 3 show an affinity to each other's bills and an aversion to the bills of members 4 and 5. The affinity and aversion in this case are reciprocated. Now to solve this matrix, we must ask only whether the vote matrix of the upper left corner has one or more rows of Y's. If it has only one such row (say, A_2), then this version of the bill will dominate the matrix. Suppose, however, the vote matrix of this partition to be: (where the Y subscripts indicate ordinal ranking)

Issue	Members 1	2	3
A_1	Y_1	Y_2	Y_3
A_2	Y_2	Y_1	Y_2
A_3	Y_3	Y_3	Y_1

Pure bargaining appears to be indicated here, unless these three members belong to one party, and a party caucus under established rules of selection (majority vote, for example) is used to determine the outcome.

We cannot ignore members 4 and 5, however. If they are prepared to vote Y on their third choice, then the game is up and A_3 will dominate the matrix in the absence of pressures external to our consideration (party loyalty, for example). Is A_3 the "right" choice?

To examine that question, look at the total combined vote and ordinal matrix as specified by our assumptions on voting:

Issue	Members 1	2	3	4	5
A_1	Y_1	Y_2	Y_3	N_5	N_4
A_2	Y_2	Y_1	Y_2	N_4	N_5
A_3	Y_3	Y_3	Y_1	Y_3	Y_3
A_4	N_5	N_4	N_5	Y_1	Y_2
A_5	N_4	N_5	N_4	Y_2	Y_1

First, although A_3 is unique in having a unanimous Y vote, that fact is not significant. Suppose members 1 and 2 vote N on their third choice; A_3 still dominates the matrix. The crux of the matter is that only member 3 is in all minimum winning coalitions and he chooses A_3.

Still delaying an answer to the question, is A_3 the right choice, let us explore the general solution of these matrices with mutually exclusive alternatives through a series of logical statements:

1) Any ordinal matrix of n voters which displays different versions of one bill A

can be reduced to an nxn matrix if there are n different first choices.

2) If there are fewer than n first choices, the matrix can be reduced to a kxn matrix where $k<n$, where $k=1, \ldots n-1$.

3) Any such nxn and kxn ordinal matrix will have vote matrices which can be combined with it. Using majority vote as the decision rule, only rows containing at least $(n+1)/2$ (if n is odd) or $(n/2)+1$ (if n is even) Y votes need be considered.

(If no row has a majority of Y votes, there is no version of a bill on the issue A which can be passed by the legislature.)

(If only one row has a majority of Y votes, then this is the only version which can be passed.)

4) If two or more rows have (at least) a majority of Y votes, selection among them takes the following form:

(a) for kxn matrices (every row passing),

(i) any minimum winning coalition (*MWC*) composed *only* of first choices is dominant (there can be no more than one such coalition in any ordinal matrix).

(ii) if no dominant row exists, then any rows with one or more first choices in an *MWC* should be compared. If there are common members of these coalitions, the common members will determine the solution. If the ordinal matrix of such common members, when reordered so as to put those members' highest preference on the main diagonal, results in a symmetrical matrix, the solution may be indeterminate.

(b) for nxn matrices (every row passing),

(i) partition the ordinal matrix to include only members who are in two or more *MWC*s. If only one such member exists, his choice dominates.

(ii) if two or more such members exist, reorder the ordinal matrix to put their highest preference on the main diagonal. If the resulting ordinal matrix is symmetrical, the solution may be indeterminate.

(iii) if the resulting matrix is not symmetrical, the common members (by bargaining, caucus vote, or whatever) dominate the nxn matrix.

It is worth noting that ordinal symmetry in either the kxn or nxn matrix denotes a possible cyclical (indeterminate) case. Here, however, the meaning of the cycle is clear and its lack of decision benign. It denotes a lack of minimum agreement on an issue and hence chooses, correctly, to pass nothing. The issue is excluded from resolution pending a new set of legislators or a reformulation of the issue which can attract a better clustering of interests.

The same four logical statements can be used to describe the actions of a majority party in a legislature (if one assumed a high degree of party discipline) or of a committee system or dominant coalition of any kind. To take any number smaller than n, however, upsets the notion of majority rule; a notion which both Rae and Taylor have shown to have a strong claim on our rational interests. Thus, such decisions have some claim to be the "right" ones, and the presence of restrictions (committee dominance, for example) which exclude some members from the decision on how an issue is framed can be suspected of turning up with "wrong" decisions even though such exclusion is a way to mask indeterminate (cyclical) matrices.

Forming issues in an election, as opposed to formulating a bill in a legislature, is a far more imprecise process. A candidate is, properly, less concerned with each issue than in the design of a package of issues and a stance of each which will win over his opponent's package. The theoretical work of Otto Davis, Melvin Hinich, and Peter Ordeshook is particularly useful in defining candidate strategies

and social welfare under various assumed distributions.

A related but separate question which has interested many writers, notably Black (1963) and Robin Farquharson, is the order in which votes are taken on bills which are interrelated. Unless one accepts party leadership as a guide both for candidate strategy and for legislative scheduling of voting, both areas can fall into indeterminancy under certain patterns of preferences.

V. Conclusions

Dennis Mueller concisely expressed the concern of many economists about the efficacy of vote-trading,

> . . . when voters are able to make and keep vote-trading agreements, their welfare will be greater than if no agreements were made. On the other hand, the resulting set of policy decisions will fall far short of being in any sense socially optimal. If the number of voters is not so large as to preclude the formation of partially stable coalitions, it is too small to remove completely the monopsony power a voter will be able to enjoy over any issue of vital importance to him. [p. 1310]

We have explored the role of party in the formation of proto-coalitions and vote-trading as the device for producing stability of outcomes while avoiding a coalition dominant on all issues. In so doing we reemphasize Madison's point (in Federalist Paper No. 10) regarding representative government as a defense against tyranny of the majority.

The monopsony problem is, however, a significant issue in representative government. Cases can be constructed in which, in a single legislature, one legislator with a strong interest in one bill can trade off many other votes to produce a majority for his bill. Notice, however, that to do so he must be in the number of minimum winning coalitions equal to the number of votes he needs on his bill. This is not an inconsiderable constraint, in theory or in practice.

An additional protection from monopsony power is the bicameral legislature—if the districts of the two houses are correctly drawn relative to one another. The point here is not "one man, one vote" since the whole of utility analysis is based on this principle,[20] rather the point is that district lines must be drawn so that representative patterns are significantly different in the two houses. What is advocated strongly by the lower house representative of district A may be safely resisted by the upper house senator whose constituency includes districts A, B, C, and D. Should a majority of these districts be of the same mind as district A, is not the senator then an advocate also? He is indeed, but if the number of senators is sufficiently restricted (which, in most senates it is not), this event happens only when a considerable portion of the state's electorate is of this mind. The case is not then one of monopsony power, but simply represents an intense minority preference which in utility terms may be accommodated through vote-trading. A further check occurs through executive veto power; the executive being the only representative of the whole electorate. These matters are discussed further in Buchanan and Tullock (1962), while Shapley has explored the theoretical problem in his formulation of compound-simple games.

There are many peculiar "legislatures," special districts, commissions, and so forth, which exist as single houses without the check of a second house or of executive veto. The political scientist's intuitive mistrust of the use of these devices for public decision making is (or should be)

[20] Note the work of J. R. Pole. It should be remembered that senates were traditionally designed to counter the individual utilities reflected in the lower houses.

rooted in the utility defect which these bodies possess. The defect occurs not only because they are prone to single dominant majorities but also because the two-party system does not operate in them. The common practice of having the Governor, or his appointee, sit on interstate agencies, for example, may be "political" but it has no connection to the utility concerns under discussion here. This practice grew up because it was administratively convenient and the issues, in the past, not so socially significant. To continue the practice now in agencies like interstate water resources commissions, or port authorities, when the decisions made by such agencies are social choices of the most critical nature, is not to be countenanced on any utility principle. Committee rule. seniority, and other twentieth century habits of general legislatures may be condemned on the same grounds. In the best of legislatures, however, it is difficult to conceive of a perfectly competitive vote market, with marginal utilities proportional to "prices." The use of the term price has no meaning. Perhaps all that can be said is that trades take place at the margin for each person, i.e., that any legislator will prefer paying a lower price (changing his vote on an item of lesser interest to him) rather than a higher price but will always be willing to trade so long as the ratio of gain to loss is above unity. While I have elsewhere demonstrated that the probability of trading increases as the number of independent issues increases, it does not follow that prices therefore tend to approach marginal conditions, for there remains no comparability among utilities.

REFERENCES

K. J. Arrow, *Social Choice and Individual Values*, 1951, 2d. ed., New York 1963.

———, "The Organization of Economic Activity: Issues Pertinent to the Choice of Market versus Non-Market Allocation," in *The Analysis and Evaluation of Public Expenditures: The PPB System*, Subcommittee on Economy in Government of the Joint Economic Committee of U.S. Congress, Washington 1969, pp. 47–63.

R. U. Ayres and A. V. Kneese, "Production, Consumption and Externalities," *Amer. Econ. Rev.*, June 1969, *59*, 284–97.

W. Bagehot, *The English Constitution*, 2d ed., London 1905.

D. S. Bell, "A Summary of the Chairman," *Daedalus*, summer 1967, *96*, 698–704, 975–77.

D. Black, *The Theory of Committees and Elections*, Cambridge 1963.

———, "Lewis Carroll and the Theory of Games," *Amer. Econ. Rev. Proc.*, May 1969, *59*, 206–10.

———, "On Arrow's Impossibility Theorem," *J. Law Econ.*, Oct. 1969, 227–47.

J. M. Buchanan and G. Tullock, *The Calculus of Consent: Logical Foundations of Constitutional Democracy*, Ann Arbor 1962.

——— and ———, "Public and Private Interaction under Reciprocal Externality," in J. Margolis, ed., *The Public Economy of Urban Communities*, Baltimore 1965.

J. S. Coleman, "The Possibility of a Social Welfare Function," *Amer. Econ. Rev.*, Dec. 1966, *56*, 1105–22.

O. Davis, M. Hinich, and P. Ordeshook, "An Expository Development of a Mathematical Model of the Electoral Process, *Amer. Polit. Sci. Rev.*, June 1970, *64*, 426–48.

A. de Tocqueville, *Democracy in America*, New York 1947.

A. Downs, *An Economic Theory of Democracy*, New York 1957.

R. Farquharson, *Theory of Voting*, New Haven 1969.

L. K. Frank, "The Need for a New Political Theory," *Daedalus*, summer 1967, *96*, 809–16.

E. T. Haefele, "Coalitions, Minority Representation, and Vote-Trading Probabilities," *Publ. Choice*, spring 1970, *8*, 75–90.

M. Hinich and P. Ordeshook, "Plurality Maximization vs. Vote Maximization: A Spatial Analysis with variable Participa-

tion," *Amer. Polit. Sci. Rev.*, Sept. 1970, *64*, 772–96.

J. E. Jackson, "A Statistical Model of United States Senators' Voting Behavior," doctoral dissertation, Harvard Univ. 1968.

G. Kramer, "Short-Term Fluctuations in U.S. Voting Behavior: An Econometric Model," mimeo 1967.

D. C. Mueller, "The Possibility of a Social Welfare Function: A Comment," *Amer. Econ. Rev.*, Dec. 1967, *57*, 1304–11.

J. R. Pole, *Political Representation in England and the Origins of the American Republic*, New York 1966.

D. W. Rae, "Decision Rules and Individual Values in Constitutional Choice," *Amer. Polit. Sci. Rev.*, Mar. 1969, *58*, 40–56.

W. H. Riker, *The Theory of Political Coalitions*, New Haven 1962.

P. A. Samuelson, "Arrow's Mathematical Politics," in S. Hook, ed., *Human Values and Economic Policy*, New York 1967, 41–51.

A. Schick, "Systems Politics and Systems Budgeting," *Publ. Admin. Rev.*, Mar./Apr. 1969, *29*, 137–51.

L. S. Shapley, *Compound Simple Games*, Santa Monica 1967.

——— and M. Shubik, "On the Core of an Economic System with Externalities," *Amer. Econ. Rev.*, Sept. 1969, *59*, 678–84.

R. C. Silva, "Relation of Representation and the Party System to the Number of Seats Apportioned to a Legislative District," *Western Polit. Quart.*, 1964, *17*, 742–69.

P. B. Simpson, "On Defining Areas of Voter Choice: Professor Tullock on Stable Voting," *Quart. J. Econ.*, Aug. 1969, *83*, 478–90.

M. Taylor, "Proof of a Theorem on Majority Rule," *Behav. Sci.*, May 1969, *14*, 228–31.

G. Tullock, "The General Irrelevance of the General Impossibility Theorem," *Quart. J. Econ.*, May 1967, *81*, 256–70.

R. Wilson, "An Axiomatic Model of Log-Rolling," *Amer. Econ. Rev.*, June 1969, *59*, 331–41.

International Trade and Capital Mobility

By ERNEST NADEL*

It has long been recognized that commodity movements and factor movements are, to a degree, substitutes for each other in international exchange (see Carl Iverson, Mountifort Longfield, James Meade, Bertil Ohlin, John H. Williams). Yet, until recently, the dominant theory of international trade, the Heckscher-Ohlin (H-O) model, had been rather thoroughly analyzed under the rigid assumption of the immobility of factors. Only in 1957, with the publication of Robert Mundell's important article, was capital mobility in a H-O model explored. This paper presents a fuller treatment of capital mobility in the H-O model. We discuss the model under conditions of tariffs on goods flows and taxes on capital relocations.

Nations may trade by exchanging goods or by exchanging their relatively abundant factors which produce those goods. We demonstrate that the substitutability between these two avenues of exchange continues to hold even under conditions of tariffs and taxes. The dynamics and equilibria are demonstrated, showing that a nation will pay for its imports *either* through exports of goods *or* through earnings on foreign-placed capital, *not* through *both* of these methods. Tariffs and tax rates will dictate which will occur, i.e., tariffs and taxes will be shown to affect the *pattern of trade*, not merely the quantities of commodities traded.

We also correct a hitherto general and unrecognized error. We show that the levying of a tariff does *not* necessarily generate a relative price differential (of final goods prices) between the trading countries equal to the tariff proportion. The relative price differential will often be less than the tariff proportion, and this holds even though the good is still imported into the country. Further, contrary to previous results (see Ronald Jones 1967), we show that, generally, a tariff-cum-tax levy will not result in the complete specialization of production.

I. Introduction to the Analysis—A Substitutability Theorem

In the standard H-O pure barter theory of international trade, the goods exports of one country are exchanged for the goods exports of the other. In his article, Mundell discussed the effect of capital mobility in such a model. Capital mobility is defined as a relocation of capital (machines) from one country to the other accompanied by a repatriation of foreign earnings. His analysis opened up a new area of analysis of exchange between nations. In the normal H-O model, a country exports the goods in whose production its relatively abundant factor is relatively intensively used. Mundell showed that the exact same free trade results, in terms of equilibrium price ratio, factor rewards, and consumption levels, can be achieved by a nation exporting its relatively abundant factor itself instead of exporting its final consumer good.

In this paper, we continue to consider the human factor of production, labor, to be immobile. (The immobility of all labor implies also the immobility of the locus of consumption in that owners of capital are themselves part of the labor force. Thus, the repatriation of foreign earnings can be viewed as an implication of the

* University of California, Berkeley.

assumption of labor immobility, and not as a separate, additional assumption itself.)

We will employ the two-country (H, F), two-good (steel, cotton), two-factor (capital, labor) model, with constant returns to scale in production, exactly identical production functions in each country, and no reversals of factor intensity occurring. Steel is always capital-intensive, and country H is initially endowed with a relative abundance of capital.

The consideration of capital mobility in effect turns the model into a three-good model. In the analysis which follows, these three goods are: steel (the home country's "natural" export good under conditions of capital immobility); cotton (the home country's natural import good under conditions of capital immobility); and capital services (a once-and-for-all relocation of a stock of capital being equivalent to a continuous flow of the services of that capital).

Assuming repatriation of foreign earnings, we will discuss the three basic patterns of exchange between the two nations that may occur:

Pattern 1: Country H can *import cotton*, and pay for it by *exporting steel*. No capital services need be traded. (This could be shown by production at P in Figure 1; consumption is at point C.)

Pattern 2: Country H can *import cotton*, and pay for it by *exporting capital services*. No steel need be traded. (This could be shown by production at J in Figure 1, consumption at C.)

Pattern 3: Country H can *import steel*, and pay for it by *exporting capital services*. No cotton need be traded. (This could be shown by production at N in Figure 1, consumption at C.)

Combinations of these patterns are possible, such as importing cotton and paying for it by exports of steel and capital

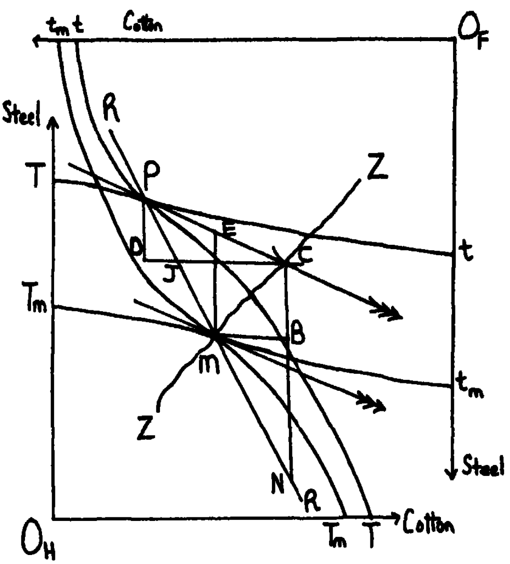


FIGURE 1—REPRESENTATION OF EQUILIBRIUM DURING FREE TRADE

services. Our analysis will deal mainly with the simple patterns discussed.

Although our model is a three-good model, it will be shown that, except for razor's edge cases, only two of the three goods will be traded internationally. And, in those razor's edge cases, (i.e., when the tariff and tax levels happen to satisfy a certain relationship), all three goods could enter international trade, but it would not be necessary that they do so. Trade in only two goods will always suffice.

When analyzing the effects of tariffs and taxes, it will be shown that it is incorrect to *assume* that the post-interference trade *pattern* will be similar to the pre-interference pattern.

We will first briefly summarize the Mundell analysis demonstrating the substitutability between goods flow exchange and factor relocation exchange between nations. Then, we will use this "substitutability theorem" (our name for the central proposition, but the analysis essentially duplicates Mundell's), to investigate fully

the consequences of interferences due to tariffs and taxes on foreign earnings.

With no domestic capital invested in the foreign country, free goods trade equilibrium will be as shown in the following diagram.

TT is country *H*'s transformation curve. Production is at *P*, consumption at *C*. Country *F*'s transformation curve is *tt*. Again, country *F*'s production and consumption are also at *P* and *C*, respectively, considered relative to the origin, O_F, of country *F*. *PD* of steel from country *H* is exchanged for *DC* of cotton from country *F*.

RR is the Rybczynski contraction (expansion) line, for the price ratio given by the slope of *PC*. At that price ratio, if there is a capital outflow (inflow), the transformation curve of *H* contracts (that of *F* expands); in so doing, at its point of intersection with *RR*, the slope of the new transformation curve is equal to the slope of *PC*. Of course, as capital moves, say, from country *H* to country *F*, the two changing transformation curves continue to be tangent to each other along *RR*, with a slope that of *PC*. Production in both countries therefore will change in a southeasterly direction. World production is unchanged.

For barter trade with factor immobility, country *H* would export *PD* of steel in return for imports *DC* of cotton. If instead exchange were to take place via a relocation of capital, a solution would, as Mundell showed, be as follows. Let *ZZ* represent the Engel curve, for price ratio given by slope *PC*, for those consumers in country *H* who do send their capital to country *F*. The curve *ZZ* crosses the expansion line *RR* at *M* which is the final production point in the two countries. The boundary *TT* contracts to T_MT_M; similarily *tt* expands to t_mt_m. Let K_M represent the implicit volume of capital relocation. Therefore *MB* of cotton plus *BC* of steel will represent the interest earnings on K_M. Furthermore, *MB* of cotton plus *BC* of steel represents the exact composition of consumption these capital exporters desire because by construction of *ZZ*, the difference between *C* and *M* represents desired consumption of capital exporters, at unchanged goods and factor prices.

Thus, factor relocation can generate the exact same results as free trade in goods with factor immobility; the only differences are that the distribution of the location of capital among countries is different and that there is a continuous (one-way) repatriation of foreign earnings, instead of a continuous (two-way) goods flow. But, in terms of factor rewards (i.e., rentals on capital, which are not the same as rates of return on capital), relative prices, and hence consumption loci, the two methods of exchange are identical. Notice, too, that *in the factor relocation case no barter exchange occurs between the two nations, in the sense that imports are paid for out of foreign earnings, and not by an equivalent amount of final goods exports.* Aside from the repatriation of foreign earnings, the two nations are self-sufficient.

(The reason why the two nations can become self-sufficient is to be found in the Rybczynski theorem. Under free trade, country *H* produces more steel and less cotton than it desires; and vice versa for country *F*. The relocation of capital causes cotton production to increase and steel production to decrease in country *H* and vice versa in country *F*. These output changes adjust so as to satisfy domestic demands so that eventually the need for imports fall to zero.)

Effects of Tariffs and Taxes

We can now proceed to analyze the effects of tariffs and taxes. As already shown, goods flow and factor relocation are substitutes. It turns out that tariff and tax interferences create incentives which will dictate, via profit maximizing behavior, whether goods flow or factor relocation will take place. These incentives

will determine which of the three patterns of exchange discussed above will actually occur. (To avoid irrelevant complications, we will assume that the government redistributes the tariff or tax proceeds back to its citizens.)

Beginning with a situation of free goods trade, we consider the effects of country H imposing a tariff (100 τ percent) on cotton imports. No tax is imposed on foreign earnings, and it is also assumed, *temporarily*, that purchases of foreign goods paid for out of repatriated earnings are *exempt* from the tariff applicable to commercial trade. As Mundell has shown, the tariff interference would raise the price of cotton in country H, so long as cotton is imported, i.e., so long as barter trade exists. And, by the Stolper-Samuelson theorem (see Wolfgang Stolper and Paul Samuelson), a higher price of cotton means a lower rental to capital. (The opposite price and factor reward movements occur in country F. The rental on capital is a unique, monotonic function of the relative goods price; because production functions are the same in both countries, this rental price function is the same in both countries.) Thus, there is an incentive for capital to move from country H to country F, i.e., capital relocation becomes preferable to goods trade. The capital outflow will cease when prices and hence rentals on capital are again equalized in the two countries. But, so long as barter trade exists, prices in the two countries will differ by the amount of the tariff. So, for rentals on capital to be equalized, barter trade must disappear; the two nations must become self-sufficient; we must reach the factor relocation solution.

One such solution (as shown by Mundell) is represented by production point M in Figure 1, which implicitly reflects the transfer of an amount of capital K_M.

Actually, if purchases of foreign goods paid for out of repatriated earnings are not subject to the tariff, any point along RR southeast of (and including) point J in Figure 1 is a possible equilibrium production point. This can be shown as follows. At unchanged terms of trade, with production at P, country H is willing to export some steel (PD) in return for some imports of cotton (DC). As capital moves from country H to country F, country H is able to purchase cotton out of its repatriated foreign earnings. Also, as the capital relocation occurs, production of steel falls and that of cotton rises in country H. Thus, excess demand for cotton falls. Eventually, as capital relocates, all the excess demand for cotton is exactly satisfied by the repatriated earnings. This occurs at point J. For a relocation of a quantity of capital equal to or greater than that implicitly represented by production point J (i.e., K_J of capital), there is no longer any need for country H to export steel to pay for desired cotton imports: i.e., there will no longer be any strict barter trade.

We now examine what occurs when the tariff on imported cotton goods also applies to the purchases of cotton goods out of repatriated earnings, i.e., when there is *no exemption*. This of course is what normally occurs when an import tariff on cotton is levied.

In this case, so long as cotton imports are coming into country H from country F, the tariff will be applied to them. If the tariff is effective, it creates a relative price differential between the two countries, hence a rental on capital differential as well. So long as this occurs, capital will continue to move from country H to country F. Capital will relocate between the two countries until there no longer are any imports of *cotton* into country H, since the tariff applies to imports of cotton only. This will occur when production point N is reached, reflecting a relocation of K_N of capital from country H to country F ($K_N > K_M$). Until production point N is reached (at unchanged terms of trade),

there will always be a desire to import cotton into country H. But, by the Rybczynski Theorem, when point N is reached, the excess demand for cotton in country H falls to zero. At that point, NC of steel are the only imports into country H, NC of steel also being equal to the rental on K_N of capital placed in country F. When point N is reached, the old free trade equilibrium results are restored.[1]

Thus, even when τ applies to all cotton imports, a capital relocation can occur which will create a successful "evasion" of the tariff. The escape valve is a change of trading pattern, with country H becoming an importer of steel, rather than a exporter of steel. There is a switch from the first pattern of trade described in Section I to the third pattern. Thus, Mundell was correct in his paper when he suggested that so long as purchases out of foreign earnings are subject to the tariff, the free trade results will be disturbed. However, he did not consider the possibility of the "switching" of trade patterns described above. Of course, with the switching possible, purchases of steel out of foreign earnings are not subject to the tariff, because there is no tariff on steel.

This evasive switch of trading pattern is rendered impossible if the tariff applies to *all* foreign goods entering the home country—*cotton or steel*; it is also impossible if a tax is levied *directly upon foreign earnings*. We now turn to analyze these cases.

II. Conditions for Capital Relocation and Correction of a Fundamental Error

We must first consider the general relationship that obtains in this model between differences in relative prices between the two countries and differences in real rentals on capital between the two countries. This is the familiar factor price/goods price relationship implicit in the technology we are assuming, and a reflection of the Stolper-Samuelson theorem. This relationship holds so long as *specialization is incomplete* in the two countries.

Because we are assuming identical technologies in the two countries, a difference in relative prices in the two countries creates (or, is accompanied by) a difference in real rentals on capital in the two countries.

We shall adopt the following symbols:

P_C = nominal price of cotton (For simplicity, we may assume that both countries use the same nominal currency standard, e.g., dollars. The nominal price would then be the dollar price.)

P_S = nominal price of steel

$p = P_s/P_C$ = relative price of steel in terms of cotton

r = real rental on capital in terms of cotton

R = nominal rental on capital $= r \cdot P_C$

dx/x = the proportional change in the variable x

Making use of the magnification effect discussed by Jones (1965), $(dr/r)/(dp/p) = \alpha > 1$. (This technological relationship is expressed in terms of the good cotton. There is no loss in generality in doing so.) Furthermore, $dR/R = dr/r + dP_C/P_C$.

We begin with a situation of free trade and no home capital placed abroad. Country H initially exports steel in return for imports of cotton.

We will consider the case where the home country levies a tariff of 100τ percent on *all* imports, cotton and steel, into country H and also levies a tax of $100t$ percent on the earnings of any capital that is placed in the foreign country.

Now, the tariff will initially raise the price of cotton in country H.

[1] It is possible that position N could not be achieved, in that it may lie off the diagram, below the horizontal axis. In this case, specialization in cotton production would occur. Position M can always be reached.

$$\frac{dp}{p}=\frac{dP_S}{P_S}-\frac{dP_C}{P_C}=-\frac{dP_C}{P_C}=-\tau$$

(Initially, the price of steel does not change, i.e., $dP_S/P_S=0$.)

This in turn leads to $dr/r=\alpha dp/p=-\alpha\tau$, and $dR/R=dr/r+dP_C/P_C=-\alpha\tau+\tau=-\tau(\alpha-1)<0$, because $\alpha>1, \tau>0$.

For ease of exposition, we hold foreign prices and foreign rental on capital constant. (This is hardly crucial at this point, for what we are actually concerned with are *differences* in prices and rentals between the two countries.)

Because of the tax ($100t$ percent) on foreign earnings, the rental that home capitalists can earn by sending their capital abroad changes by $-100t$ percent. But, the tariff causes $dR/R=-\tau(\alpha-1)$ in country H. The tariff-cum-tax combination will cause a capital outflow from country H to country F only if the fall in the nominal return to capital at home is larger than the decrease in the net return earned by home capitalists who would invest abroad;

$$\text{i.e., if} \quad -\tau(\alpha-1)<-t,$$

$$\frac{t}{\tau}<\alpha-1>0$$

Thus, we have derived the condition for an induced capital outflow consequent to levying a tariff-cum-tax combination. Capital will relocate from country H to country F if $t/\tau<\alpha-1$. (If $t/\tau=\alpha-1$, home capitalists will be indifferent as to where they place their capital. Hence here, and throughout the paper, we have used only the inequality conditions. In our example of the previous section, with $t=0$, we had $t/\tau=0/\tau<\alpha-1$, so that the tariff (-cum-tax) combination did indeed cause a capital outflow.)

We can also specify the equilibrium conditions that must hold when the capital relocation ceases. Net nominal rentals to home capitalists must be equal in the two countries. The gross foreign rental is held constant at its free trade level, by assumption. Thus, compared to free trade magnitudes, the rental to home capitalist investors abroad will change by $dR/R=t$. This must be equal to the change in the nominal rental to capital in country H.

Assume for now that only cotton is imported into country H. (Below, we will present the necessary conditions for this to occur.) Thus, $dP_C/P_C=\tau$, (i.e., the relative change in the price of cotton in country H, relative to its free trade level, and hence also relative to the unchanged price of cotton in country F, will be exactly equal to the tariff proportion.) Let dp^*/p represent the proportional change in the home relative price of steel that will have occurred once equilibrium is reached. (Thus dp^*/p also represents the final equilibrium relative price differential between the two countries.)

We again express the change in the home nominal rental:

$$\frac{dR}{R}=\frac{dr}{r}+\frac{dP_C}{P_C}=\frac{dp^*}{p}\cdot\alpha+\tau$$

In equilibrium, we must have

$$\frac{dp^*}{p}\cdot\alpha+\tau=-t,$$

i.e.,

$$\frac{dp^*}{p}=-\frac{t+\tau}{\alpha}$$

Summarizing, we have the following:

For a capital relocation to occur, we must have $t/\tau<\alpha-1$.

Once equilibrium has been restored (when the capital relocation has ceased), we must have $dp/p^*=-(t+\tau)/\alpha$.

Normally, it is assumed that a tariff of 100τ percent will cause a relative price differential in equilibrium of $dp^*/p=-\tau$. By our analysis, this implies $dp^*/p=-\tau$

$= -(t+\tau)/\alpha$. This implies $t/\tau = \alpha - 1$. However, this contradicts our conclusion that for a capital relocation to even occur we must have $t/\tau < \alpha - 1$. Clearly, an impasse has been reached.

We have come to the crux of a very important issue. If capital relocation is to occur, *and* incomplete specialization is also to be retained, *and* an equilibrium state is to be finally reached, the relative price differential will be equal to the tariff only if $t/\tau = \alpha - 1$. However, we have seen that the condition for a capital relocation to occur in the first place is $t/\tau < \alpha - 1$.

The problem can be demonstrated in another way. By the Stolper-Samuelson theorem, and given our assumptions about the production functions in the two countries, there is a unique relationship between relative price differentials and rental on capital differentials between the two countries, so long as specialization is *incomplete* in the two countries. When a country levies a tariff (100τ percent) and a tax ($100t$ percent), it seeks to generate a relative price differential of 100τ percent and a relative nominal rental on capital differential of $100t$ percent. It is only by chance that the tariff and tax levied will be related to each other in the *exact* proportion that the technology (i.e., the Stolper-Samuelson theorem) *necessitates*. If that fortunate situation does not occur, something in the system must "break down."

It was Jones (1967) who first recognized the existence of this impasse. Until his analysis, no writers[2] had worried about the constraints that technology (i.e., the Stolper-Samuelson theorem) imposes on the system when tariffs and taxes are levied. Jones' resolution of the difficulty was to state that it must be incomplete specialization which breaks down, one or both countries being forced to specialize in production. When this happens, the Stolper-Samuelson theorem no longer imposes the constraint of the magnification effect. With incomplete specialization gone, we can no longer require that in equilibrium $dp^*/p = -(t+\tau)/\alpha$.

Jones' resolution, although it does overcome the impasse, is not in general the correct one—i.e., market forces will not in general bring us to his solution. Instead, the resolution which we present not only resolves the impasse, it is also the solution which market incentives will dictate. (The Jones solution is correct, and is brought about by market forces only if steel is eventually imported into the country *and* the point N (Figure 1) lies off the diagram. However, with a uniform tariff on *all* imports, steel is never imported into the country. The Jones analysis, however, did *not* consider a *uniform* tariff levied on cotton *and steel* imports.)

We attempt to show that as a result of an induced capital relocation incomplete specialization does not break down. Instead, we show that relative goods prices in the two countries will not differ by the amount of the tariff—they will differ by *less*. This is rendered feasible because the pattern of international exchange is drastically altered. Before the capital relocation both goods, steel and cotton, enter into trade. In such an instance, relative prices must differ by the amount of the import tariff. But, because of the capital relocation, one of the final goods, steel or cotton as the case may be,[3] *no longer enters into international trade* (i.e., we change from pattern 1 of trade, to patterns 2 or 3, cf. Section 1).

Of course, the *nominal* prices of the good that remains internationally traded

[2] One can safely say that practically all writers in trade theory have implicitly assumed the joint effectiveness of any combination of tariff and tax (except in the cases of totally prohibitive tariffs or tariffs on non-imported items).

[3] This depends on the relationship between the tariffs on steel and on cotton. With a uniform tariff, steel disappears from trade. See more detailed results below.

do differ by the tariff magnitude as between countries. But relative prices do not have to differ by the magnitude of the tariff.

Essentially, the error others have made is to *assume* what the final pattern of international exchange will be like, i.e., similar to the pattern observed before capital mobility and the presence of tariffs and taxes are considered. What we show is that such an assumption is, in general, not valid. The pattern of trade must come out of the analysis—it must not be assumed. What we will show below is that the tariff and tax parameters will actually determine what that pattern will be like.

We will substantiate these propositions more fully below. At this point we will present a diagrammatic summary of our results.

In Figure 2 we have drawn a line with slope $\alpha-1$. The locus of positive t and τ that country *II* can set is the entire (first) quadrant. For any combination of t and τ lying above the line with slope $\alpha-1$, capital relocation will not occur. For any combination of t and τ lying below the line, capital relocation will occur (because $t/\tau<\alpha-1$).

Let V be one such latter point, representing a combination of a tariff of $100\tau_v$ percent and a tax of $100t_v$ percent which will lead to capital relocation. We can show what the final equilibrium relative price dp_v/p will be when the capital relocation has stopped, i.e., $dp_v/p=(t_v+\tau_v)/\alpha$.

To show dp_v/p on the figure, draw a line through V having a slope of $-45°$. That line crosses the line $t/\tau=\alpha-1$ at A. The abscissa of point A is $dp_v/p=(t_v+\tau_v)\alpha$ $(=CB=OE)$.

Also notice that $dp_v/p=OE<OD=\tau_v$, i.e., $dp_v/p<\tau_v$.

We can also show another interesting result. We can show that the change in the *real* return to capital is *greater* than $100t$ percent. If we measure the change in the

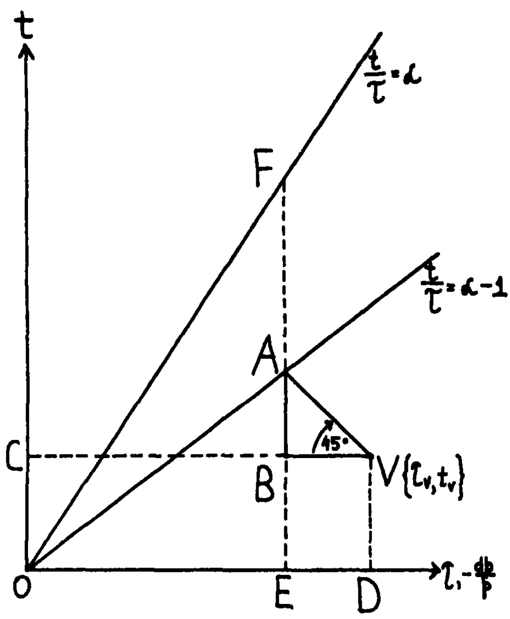


FIGURE 2 · EQUILIBRIUM MAGNITUDES AFTER CAPITAL RELOCATION OCCURS

real return to capital in terms of cotton,

$$\left(\frac{dr_c}{r_c}\right) = \alpha\left(\frac{dp}{p}\right)_v = EF \text{ (on the figure)} > t_v$$

Thus, if capitalists consume only cotton, their real return falls by an amount greater than the tax on foreign earnings. And, even if they consume only steel, we get

$$\frac{dr_s}{r_s} = (\alpha-1)\left(\frac{dp}{p}\right)_v = EA > t_v$$

Thus, no matter what their consumption pattern is like, the real return to capital will fall as a result of the tax-cum-tariff combination, and it will fall by *more* than the rate of tax on foreign earnings.

III. Equilibrium After Factor Relocation

Algebraic Presentation

Before we derive any further results we must demonstrate that what we said happens actually does happen, i.e., we must show that, when $t/\tau<\alpha-1$, incomplete specialization does not disappear,

but instead the induced capital relocation causes a change in the trading pattern, country H no longer exporting steel, and the relative price differential between the two countries being less than the tariff.

We begin with a situation of free goods trade, and no initial capital relocation. Production is represented by P in Figure 1. Country H imports DC of cotton and exports DP of steel. Now country H levies a tariff on all goods imports of 100τ percent, and a tax on the earnings of capital placed in the foreign country of $100t$ percent. All foreign earnings are repatriated.

If $t/\tau > \alpha - 1$, no capital relocation occurs. The tariff raises the price of cotton in country H by 100τ percent, but the difference in nominal returns to capital between the two countries is less than $100t$ percent.

If $t/\tau < \alpha - 1$, capital relocation occurs. The price of cotton in H rises by 100τ percent. By the Rybczynski theorem, because of the capital relocation, production of cotton rises and that of steel falls in country H. The opposite occurs in country F. The excess demand for cotton to be imported into country H falls. Country H now pays for its imports of cotton partly through its earnings on foreign capital, partly through its exports of steel. As capital continues to relocate, exports of steel continue to fall. Eventually, exports of steel are zero, and the value of cotton imports is exactly equal to the value of earnings on foreign placed capital. At that point, we have the price of cotton 100τ percent higher in country H than in country F and the price of steel equal in the two countries (because country H was exporting steel). But there still is an incentive for more capital to move to country F (i.e., the capital market is still not in equilibrium). As this relocation continues the production of steel falls further in country H, thus giving rise to a demand to *import* steel into country H; i.e., the price of steel begins to rise. However, steel will not be imported into country H until the price of steel there rises by 100τ percent because there is a tariff on all imports, steel as well as cotton.

As the price of steel rises in country H, the relative price of steel in terms of cotton rises too. This raises the real and also the nominal rental to capital in country H. This reduces the incentive for capital to move to country F. As soon as the price of steel rises sufficiently to make the relative prices in the two countries differ only by $(t+\tau)/\alpha < \tau$, the capital relocation ceases. At that point, the net nominal returns to home-owned capital are equal in the two countries. No more capital leaves country H. We have $(t+\tau)/\alpha < \tau$ because we began with a situation where $t/\tau < \alpha - 1$.)

We can see how the pattern of trade has changed. Country H now imports cotton, but pays for it through capital exports, not steel exports. The nominal prices of cotton differ by 100τ percent between the two countries. The nominal prices of steel differ by *less* than 100τ percent, the price of steel being higher in country H. Nevertheless, steel is not imported into country H because of the tariff on all imports.

Actually, the tariff and tax incentives create a situation analogous to that of capital immobility—except that now *steel* is in a sense immobile (not by assumption, but by market forces) instead of capital.

Before we present a diagrammatic illustration of the equilibrium after factor relocation in the presence of tariffs and taxes, we can derive some more algebraic results.

We can analyze the impact of *differential* tariffs on cotton imports and steel imports. Let these be represented by τ_C and τ_S. Let $100t$ percent again be the tax on foreign earnings. We can demonstrate a relationship between τ_C, τ_S and t which will determine, under conditions of capital

relocation, whether country H, in final equilibrium, will be importing cotton or steel and which good will disappear from international trade; i.e., we can show the conditions necessary for pattern 2 or pattern 3 type of exchange to occur (see Section I).

We saw above, that for $\tau_C=\tau_S>0$, $t>0$, the equilibrium differential in relative prices was $dp/p=-(t+\tau)/\alpha$. Now, $dp/p=dP_S/P_S-dP_C/P_C=-(t+\tau)/\alpha$. For $\tau_S=\tau$, we had $dP_C/P_C=\tau$.

$$\therefore\ \frac{dP_S}{P_S}=\frac{dp}{p}+\frac{dP_C}{P_C}=-\left(\frac{t+\tau}{\alpha}\right)+\tau$$
$$=\frac{\tau(\alpha-1)-t}{\alpha}$$

But, capital relocation occurs only if $t/\tau<\alpha-1$, i.e. $\tau(\alpha-1)>t$. Therefore, $dP_S/P_S=[\tau(\alpha-1)-t]/\alpha>0$; also, $dP_S/P_S=\tau-\tau/\alpha-t/\tau=\tau-(\tau+t)/\alpha<\tau$, so $\tau>dP_S/P_S>0$.

Thus, if $\tau_S=\tau_C=\tau$, steel is not imported into the home country, because the domestic price of steel is less than $100\tau_S$ percent greater than the foreign price. Steel would be imported if τ_S were lowered to the point where τ_S was less than the difference between home and foreign prices, i.e., if $\tau_S<[\tau_C(\alpha-1)-t]/\alpha$.

Thus, for a situation where capital relocation occurs steel will be the only good imported into country H if

$$\tau_s<[\tau_C(\alpha-1)-t]/\alpha$$

What this means in terms of the capital relocation process described above is as follows. Initially, country H imports cotton and exports steel. After the tariffs and tax are levied, capital relocates to country F. Eventually, the point is reached, because of the Rybczynski effects, where no steel is exported. Cotton imports are paid for totally out of earnings on foreign-placed capital. But the capital relocation continues beyond that point, because the capital market is not yet in equilibrium, giving rise to an excess demand for steel. The price of steel rises in country H. It can rise no higher than $(1+\tau_S)P_S^F$. When P_S^H reaches $(1+\tau_S)P_S^F$, steel starts being imported into country H, as well as cotton—both imports are paid for out of capital's foreign earnings. But, the capital relocation does not cease, because relative prices differ between the countries by an amount sufficient to retain the incentive for capital relocation.

Again, by the Rybczynski theorem, the capital relocation causes an increase in production of cotton in country H. This means that as capital moves to country F, desired cotton imports fall. Eventually these desired cotton imports fall to zero, domestic production of cotton in H having expanded sufficiently to satisfy domestic demand. But capital continues to relocate beyond the point where cotton imports are zero, because relative prices haven't changed. As the capital continues to move, cotton production increases in country H. This causes an excess supply of cotton, which lowers the price of cotton to $P_C^H<(1+\tau_C)P_C^F$. When P_C^H falls low enough so that the relative price differential between the countries has narrowed so as to cut off the incentive for capital relocation, capital relocation ceases. And, at that final equilibrium, steel is the only good being imported into country H.

We can now summarize our algebraic results:

A capital relocation will occur if $t/\tau_C<\alpha-1$.

If $t/\tau_C<\alpha-1$, cotton will be the only good imported into country H if $\tau_S>[\tau_C(\alpha-1)-t]/\alpha$, and the equilibrium production point will be "in the vicinity" of J in Figure 1. Furthermore, we will have $dp^*/p=-(t+\tau_C)/\alpha>-\tau_C$: also, $\tau_s>dP_S/P_S=[\tau_C(\alpha-1)-t]/\alpha>0$.

If $t/\tau_C<\alpha-1$, and $\tau_S<[\tau_C(\alpha-1)-t]/\alpha$, steel will be the only good imported

into country H, and the equilibrium production point will be "in the vicinity" of N in Figure 1. Furthermore, we will have $dp/p^* = -[\tau_S + t]/(\alpha - 1)$; also, $dP_C/P_C = [t + \alpha\tau_S]/(\alpha - 1) < \tau_C$.

Notice four more particular results:

If $\tau_C = t$, a capital relocation occurs only if $\alpha > 2$.

If there is an undifferentiated tariff as between cotton and steel (i.e., $\tau_S = \tau_C$), and if $t/\tau_C < \alpha - 1$, cotton will be the only imported good, because $\tau_S > [\tau_C(\alpha - 1) - t]/\alpha$.

If $\tau_S = 0$, and $t/\tau_C < \alpha - 1$, steel will be the only imported good because $\tau_S < [\tau_C(\alpha - 1) - t]/\alpha$.

Even if $t = 0$, $\tau_S > 0$, $\tau_C > 0$, capital relocation will occur but the real rental on capital will still fall in country H.

Geometric Presentation

We must now show what the equilibrium looks like, diagrammatically, for a combination of τ and t such as is represented by V in Figure 2. With a tariff τ_v (uniformly applied to steel and cotton imports) and a tax on foreign earnings t_v, we know from our previous analysis that the final equilibrium price differential dp/p between the two countries will be $dp^*/dp = (t_v + \tau_v)/\alpha = W$.

In Figure 3, we show the equilibrium for country H; P is the production point during free goods trade; C is the consumption point. (We hold the foreign price ratio constant, for ease of presentation.)

If we were to consider the case of a tariff of $W = -dp^*/dp = (t_v + \tau_v)/\alpha$, with capital immobility, we would end up with production at P' where the slope of the transformation curve at P' is $(1+W)$ times the slope at P; consumption would be at C'; EE, which runs through P' is parallel to PC, and its slope is equal to the foreign price ratio. At the point C' a domestic community indifference curve crosses EE and has a slope of $(1+W)$ times the slope of EE.

The solution in our capital relocation case is very similar. We know that the domestic price ratio will differ by W from the foreign price ratio. So, through P' we draw the Rybczynski line $R'R'$. Domestic production will be on $R'R'$, southeast of P'. Each point on $R'R'$ represents some quantity of capital outflow; that capital will earn returns abroad, which will be taken home in the form of cotton. These returns must be added to domestic production to get domestic production plus foreign cotton earnings. At C', where EE cuts an indifference curve with a slope equal to $(1+W)$ times the foreign price ratio, we get the equilibrium consumption point; to find the production point, we draw a horizontal line through C', which cuts $R'R'$ at Y. The production point Y is on a new contracted transformation curve $T'T'$.

If there were differential tariffs set (τ_S, τ_C), such that $\tau_S < [\tau_C(\alpha - 1) - t]/\alpha$, so that capital relocation occurred and only steel

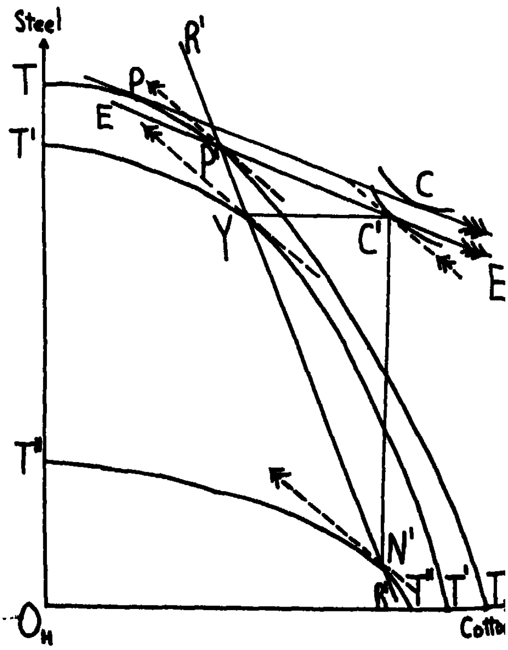


FIGURE 3—EQUILIBRIUM WITH TARIFFS AND TAXES

were imported into country H, equilibrium would be at N' in Figure 3, i.e., at a point on $R'R'$ vertically below C'. It is possible that point N' could not be reached, i.e., that country H might end up specializing in cotton. Point Y can always be reached, i.e., there is no chance of specialization occurring if equilibrium is to be at Y.

That the EE line should have the exact same slope as the foreign price ratio is not at all random. It is dictated by the technology. It is a reflection of the duality of simple general equilibrium models, of the intimate relationship between the Rybczynski and Stolper-Samuelson theorems.

The algebraic proof of this is somewhat arduous. It can be shown intuitively as follows. $P'C'$ is the home country's budget constraint, before capital relocation, expressed in terms of foreign prices. The capital relocation merely causes identically opposite production changes in both countries, leaving the foreign valuation (i.e., in terms of foreign prices) of domestic income unchanged.

Thus, any position of final consumption (utility) achieved through restrictions on barter-goods trade can be achieved through restrictions on factor-relocation trade. (Furthermore, government revenue is exactly the same in the two cases.) The "substitutability" between the two continues to hold. The introduction of factor mobility does not increase the optimization potential a nation has when it seeks to interfere with barter goods trade under conditions of capital immobility. What capital mobility does is provide a potential method of frustrating the intended results of attempted interferences in the barter trade flows.

REFERENCES

C. Iversen, *Aspects of the Theory of International Capital Movements,* Boston 1935, ch. 2.

R. W. Jones, "The Structure of Simple General Equilibrium Models," *J. Polit. Econ.*, Dec. 1965, *73*, 557–72.

———, "International Capital Movements and the Theory of Tariffs and Trade," *Quart. J. Econ.*, Feb. 1967, *81*, 1–39.

M. Longfield, *Lectures on Political Economy,* Dublin 1834.

J. E. Meade, *Trade and Welfare,* London 1955.

R. A. Mundell, "International Trade and Factor Mobility," *Amer. Econ. Rev.*, June 1957, *47*, 321–35.

B. Ohlin, *Interregional and International Trade,* Cambridge, Mass. 1967.

T. M. Rybczynski, "Factor Endowment and Relative Commodity Prices," *Economics,* Nov. 1955, *22*, 336–41.

P. A. Samuelson, "Prices of Factors and Goods in General Equilibrium," *Rev. Econ. Stud.*, 1953, *21*, no. 1, 1–20.

W. R. Stolper and P. A. Samuelson, "Protection and Real Wages," *Rev. Econ. Stud.* Nov. 1941, *9*, 58–73.

J. H. Williams, "The Theory of International Trade Reconsidered," *Econ. J.*, June 1929, *39*, 195–209.

Peasants, Procreation, and Pensions

By Philip A. Neher*

The neoclassical theory of population employs cost-benefit analysis to predict the size of family which fertility decision makers desire. While the theory has not been articulated in detail, its broad outlines have been made clear enough to guide the formulation of testable hypotheses.[1] Roughly put, the theory treats children as producers' and consumers' durable goods, yielding a stream of benefits which are compared with a stream of costs. Births occur if the present value of the benefits exceeds the present value of the costs.

The theory has been used by Malthusians to rationalize the casual observation that population growth rates increase during the early stages of economic development and then, perhaps, fall later on as even higher levels of per capita income are attained. The implications of this behavior have been explored in dynamic as well as static macroeconomic models.[2]

As useful as these models have been in improving our understanding of growth and development processes, it must be said that the underlying Malthusian theory of population lacks precisely formulated micro-foundations. These foundations are important for familiar reasons. First, they are required to sharpen the predictive accuracy of the theory itself. Second, they must be secured before effective population control schemes can be devised.

Until recently, the population control effort has focused on devising inexpensive and morally acceptable birth control measures. Yet the demographic history of France and Ireland suggests that these measures are not necessary conditions for population control.[3] They are probably not sufficient conditions either. People must *want* to reduce *desired* family size. Modern birth control measures simply make it easy to do.[4]

If parents believe they can make themselves better off by having large families, they will do so. If economic considerations impinge on fertility decisions, public policy measures designed to lower fertility rates are likely to fail unless they strike directly at basic economic motives for having children. Thus, propaganda and pills alone are unlikely to have a major impact on reducing fertility if it is clear to family planners that they can improve their condition by having more children.

A major motive for having children in primitive societies is the *pension motive*. Parents invest in their children by bearing their rearing costs in anticipation of retirement when their children, in turn, will support them. I supress other motives for having children and thereby sacrifice a

* Associate professor of economics, University of British Columbia. I am grateful for discussions with Keizo Nagatani, John Cragg, and Anthony Scott, and for comments by an anonymous referee.

[1] A widely available statement of the theory is in Milton Friedman (pp. 207–11). Harvey Leibenstein (pp. 159–70), outlines the theory in the context of economic development. Examples of empirical applications are found in Gary Becker and Theodore Shultz.

[2] John Buttrick (1958) has made population growth endogenous in Robert Solow's model. Richard Nelson used a Malthusian approach in formulating his well known "low-level equilibrium trap." Jürg Niehans extends Nelson's model to allow for capital accumulation. John Conlisk explores a fascinating model which is related to Niehan's. Buttrick's (1960) contribution is similar to Nelson's and is an excellent exposition of the Malthusian model. Eric Davis and Robert Merton have explored the implications of Malthusian population growth for the optimal time path of saving in a growing economy.

[3] For other examples, see E. A. Wrigley.

[4] Leibeinstein is a major proponent of this view (see p. 159).

large measure of realism in an effort to gain precision. I shall make a number of simplifying assumptions in addition as I go along—some of them are quite drastic. What emerges is an economic theory of population in a primitive economy which, however abstract and unrealistic, seems to illuminate fundamental dimensions of human fertility in societies where children commonly provide for their parents during their retirement years. The model is thoroughly neoclassical; no new concepts are involved. I am trying to forge a new tool of analysis, but with off-the-shelf components.

The pension motive for having children will be examined in the context of a primitive economy and a natural life cycle of dependent childhood, productive parenthood, and dependent grandparenthood. The life cycle gives rise to an optimal stock of children from the parents' point of view and they control their fertility accordingly. The optimal stock of children is associated with an optimal lifetime consumption profile which is calculated on the assumption that future births are exogenous events.

I shall be analyzing a primitive economy where the only property is land, and the rights to it are vested in families or extended families. People do not, as individuals, have a claim on property income, nor do they have a claim on the fruits of their own labor. Income is distributed in accordance with a *share alike ethic* whereby all members of the family have equal claim to the product whether they work or not.

I assume further that goods cannot be carried forward through time. Parents cannot provide for their unproductive retirement years by storing goods. Thus, the *only* way for parents to consume while retired is by means of an income transfer from the working generation. This transfer is assured by the share alike ethic which serves as an unwritten and continuing intergeneration contract whereby the workers agree to support the retired.

If fertility were exogenous, and a constant force, there would be no decisions to make. Life cycles would simply repeat themselves as exact replicas of one another. But fertility *is* a decision variable in the model which I shall set out in the next section. People *can* make themselves, and future generations, better or worse off, by controlling their fertility. What *will* they do?

How do egoism and shortsightedness bear on the fertility decision?

What is the consequent optimum size of family, in equilibrium, from the point of view of the independent decision maker? Is there "overpopulation" in some sense, and is it significant due to the pension motive alone?

How might feasible institutional reform impinge upon the fertility decision? Would the introduction of alternate sources of pensions significantly reduce the pension motive for fertility?

These questions are explored in subsequent sections, followed by some highly tentative conclusions.

I. The Model

Imagine a continuing (extended) family unit composed of three equally spaced generations: grandparents (G) who have retired and perform no economic functions; parents (F) who work a fixed amount of land;[5] and children (S) who, like the grandparents, are dependent on the working parents. The total population (P) equals $S+F+G$. The passage of a unit of time turns children into parents, parents into grandparents, and grandparents into their graves. Simultaneously, a new generation of children appears.

I assume that the fertility decision is made and effected by the parents at the instant they move into the parent generation. The parents thus look forward to one full period during which they work and

[5] One can think of land as a composite, non-augmentable non-labor input.

their children are dependent, and one full period during which their children work and they (i.e., the parents) are dependent. These two periods are followed by a third during which the children of the original parents are dependent grandparents.

The consequences of a fertility decision thus span three periods, but the parents are dead during the third. Nevertheless, parents may have some concern for the state of the world after they are dead, while their children are still alive. To allow for this possibility, I designate a utility function which embraces the three periods during which the children will live. The parents' egoistic utility depends upon their own consumptions (and, because of the share alike constraint, everyone else's too) during the first two periods. The parent's concern for the future is represented by the utility of consumption enjoyed by their children (and everyone else) during the third period. I assume that utility is additive, so the parents choose to have the number of children (S_1) which will maximize[6] the following utility function, where D and E represent discount factors:

$$(1) \qquad U = u(c_1) + D\cdot u(c_2) + E\cdot u(c_3); \quad u'(c) = \mu > 0; \quad u''(c) = \mu' < 0$$

The constraints are production conditions, the share alike ethic, the fertility of future generations and interperiod mortality.

The production relation, $X=f(F)$, has the traditional form with first increasing and then decreasing marginal returns to labor as more labor is added to the land. It is illustrated in Figure 1. Note that the marginal product of labor (w) equals the average product of labor (X/F) where labor's average product is a maximum.

The share alike ethic has the obvious implication that the consumption effects of fertility behavior are felt equally by everybody, not just by the parents.

The fertility of future generations is in the hands of future parents and exogenous to the current period decision makers.[7]

Interperiod mortality is represented by p, the probability that a child (S) will live long enough to become a parent (F) and by q, the probability that a parent (F) will live long enough to become a grandparent (G). For example, $F_2=pS_1$ and $G_3=qF_2=pqS_1$, where the subscripts refer to periods.

These constraints are summarized in equations (2) through (4). Fixed and exogenous variables are denoted by a superscript bar.

$$(2) \qquad c_1 = \frac{\overline{X}_1}{P_1} = \frac{f(\overline{F}_1)}{S_1 + \overline{F}_1 + \overline{G}_1}$$

$$(3) \qquad c_2 = \frac{X_2}{P_2} = \frac{f(pS_1)}{\overline{S}_2 + pS_1 + q\overline{F}_1}$$

$$(4) \qquad c_3 = \frac{\overline{X}_3}{P_3} = \frac{f(p\overline{S}_2)}{\overline{S}_3 + p\overline{S}_2 + pqS_1}$$

The share alike ethic compels the first period product ($\overline{X}_1$), produced by the existing parents ($\overline{F}_1$'s), to be shared alike with the retired grandparents ($\overline{G}_1$'s) and however many children (S_1's) are chosen. First-period consumption (c_1) falls as S_1 rises.

$$(5) \qquad \frac{dc_1}{dS_1} = -\frac{1}{P_1}c_1$$

[6] To preserve simplicity, I assume that S_1 is a continuous and unbounded control variable. The skeptical reader can set up the problem with S_1 as a discontinuous and bounded control. The results are substantially the same.

[7] There seems no very good way to make future fertility endogenous. One possibility is the game-theoretic approach. See Edmund Phelps and R. A. Pollak. But that is another exercise.

The first-period consumption consequences of fertility are illustrated by the dotted c_1 line in Figure 1. Starting with a point like M (ignore the c^* line for now), maximum c_1 is obtained by having no children ($S_1=0$); having more reduces per capita product.

Second-period per capita consumption increases as more workers are added so long as labor's marginal product, w, exceeds per capita consumption, $c=X/P$, (not per worker product, X/F).

$$\frac{dc_2}{dS_1} = \frac{p}{P_2}(w_2 - c_2) \tag{6}$$

Starting at point M, second-period consumption, denoted by the dotted c_2 line, is zero if $S_1=0$. It rises with increased fertility up to point N where the marginal contribution of an additional worker equals average consumption ($w_2=c_2$), and then falls.

Third-period per capita consumption falls with larger numbers of first-period children. A given amount of third-period product yields less per capita consumption if it must be shared with a larger retired generation.

$$\frac{dc_3}{dS_1} = -\frac{pq}{P_3}c_3 \tag{7}$$

A necessary condition for a maximum U with respect to S_1 is that dU/dS_1 vanish.

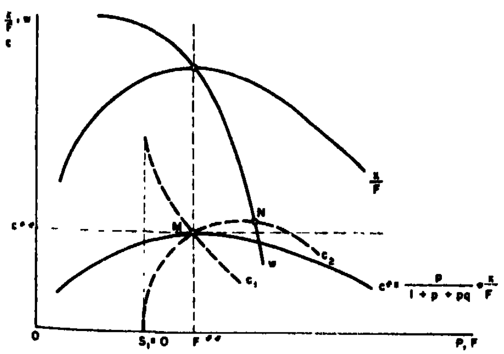


FIGURE 1

Using (5), (6) and (7), along with (1), the condition is

$$\hat{w}_2 = \frac{1}{pD}\left(\frac{\mu_1}{\mu_2}\cdot\frac{P_2}{P_1}\hat{c}_1 + Dp\hat{c}_2 + E\frac{\mu_3}{\mu_2}\cdot\frac{P_2}{P_3}pq\hat{c}_3\right) \tag{8}$$

where the superscript hat denotes a value associated with utility maximizing by the parents. Equation (8) is the condition for *economic* equilibrium.

The expression appears less formidable when applied to a steady state, or Golden Age,[8] marked by a stable and stationary population.[9] These conditions imply constant product, steady per capita incomes (and utilities) and an unchanging marginal product of labor (w). For a Golden Age, (8) is written

$$\hat{w}^* = \frac{1}{pD}(1 + pD + pqE)\hat{c}^* \tag{8'}$$

where the asterisk denotes a Golden Age value.

This condition is still not readily interpreted. However, I note that consumption per capita is related to output per worker in a Golden Age by the equilibrium demographic structure of the population. Specifically, since $F^*/P^*=p/(1+p+pq)$, I can write

$$c^* = \frac{p}{1+p+pq}\cdot\left(\frac{X}{F}\right)^* \tag{9}$$

This is the c^* line in Figure 1. Note that c^* reaches a maximum value where $w=X/F$.

Using (9) and (8′), I obtain an expression relating output per worker to labor's

[8] A Golden Age is identified in modern growth literature as a balanced growth state where every variable, if it grows at all, grows at the same proportional rate. Here, a Golden Age refers to a constant population with each generation proportional in size to every other.

[9] Demographers call a population stable if its demographic structure is unchanging. A stationary population has constant numbers. See G. W. Barcley.

marginal product

$$(10) \qquad w^* = J \cdot \left(\frac{X}{F}\right)^*$$

$$J = \frac{1}{D} \cdot \frac{1 + pD + pqE}{1 + p + pq}$$

$$J_D < 0, \qquad J_E > 0$$

This equation combines conditions for economic equilibrium (the parents are satisfied with the size of the oncoming generation), and demographic equilibrium (there is a Golden Age where the population is stable and stationary).

II. The Golden Rule of Fertility[10]

Equation (10) relates the Golden Age values of w and X/F. If $J=1$, the Golden Age population enjoys the maximum sustainable level of per capita income at point M in Figure 1. There, $c^*=c^{**}$, $F^*=F^{**}$ and $P^{**}=(1+p+pq)F^{**}$. The double-star denotes a *Golden Rule* value. Each generation of parents can say "we are procreating for future generations as we would have had past generations procreate for us."

The value of J depends upon interperiod survival rates (p and q) and the discount factors assigned to future utilities (D and E). The value of J will equal unity if

$$(11) \qquad D = \frac{1 + pqE}{1 + pq}$$

Then the Golden Rule is being observed. Clearly, a sufficient condition for compliance is that $D=E=1$. This means that parents weigh equally their utilities while working and retired *and* that they have an equal regard for their children's utility. Such farsightedness ($D=1$) and altruism ($E=1$) harmonizes the self-interest of individual parents and of the continuing family.[11]

Curiously, generalized discounting of the future ($E<D<1$) can be consistent with the Golden Rule. For example, if $p=q=E=1/2$, then $J=1$ if $D=9/10$. However, there is no uniform, non-zero, rate of discount (r) consistent with the Golden Rule. If $D=1/(1+r)$ and $E=1(1+r)^2$, (11) holds only if $r=0$.

But suppose parents care not at all for their children apart from sharing alike with them while alive so that E equals zero. Suppose further that D equals one. Then

$$J = \frac{1 + p}{1 + p + pq} < 1$$

and the corresponding Golden Age will be marked by "overpopulation" with labor's marginal product less than its average product. But the degree of overpopulation may or may not be "large," depending on the nature of the production function and the magnitude of intergeneration mortality.[12]

I have been working with a utility function in which the quality of life (consumption per capita) alone counts. But it may be that children yield a flow of consumption delights for their parents and are therefore valued for themselves.[13] In that case, the maximizing behavior of parents results in overpopulation in the sense I have used it.

[10] Compare Phelps' "Golden Rule of Procreation." The concepts are related but derived in different contexts. Phelps' Rule is for a dynamic Golden Age with capital accumulation and children valued, in part, as consumption good. My Rule is for a static Golden Age where children are valued solely as an investment for pension purposes.

[11] The analogue in growth theory is well known. Altruism and farsightedness will lead private savers toward a "consumption turnpike" along which consumption per head is the sustainable maximum. See Samuelson (1965).

[12] For example, if $p=q=1/2$ then $J=6/7$. If the production function is of the form $X=L^2/(A+B\cdot L^3)$, then $L^{**}=(A/2B)^{1/3}$ and $(X/L)^{**}=(2/3)(A/2B)^{1/3}$. But if $J=6/7$, then $L^*=(8A/13B)^{1/3}$ and $X/L^*=(13/21)(8A/13B)^{1/3}$. The ratio of the latter X/L to the former is 0.995. I would judge overpopulation to be "small" in this case.

[13] This is a common assumption in the literature. See, for example, Becker (p. 211) and Phelps (p. 181).

It is left to the reader to insert children as an argument in the utility function and work out the consequences.

III. A Fundamental Market Failure

It is well known that economic inefficiency is likely to occur if the costs and benefits relating to a decision are not wholely internalized to the decision making unit. It was found in the preceding section that the Golden Rule will be violated if parents are farsighted with respect to their retirement years ($D=1$) but neglect the welfare of those who live on after the parents die ($E=0$) but who, nevertheless, are affected by the parents' fertility.

Except in societies which practice instant reincarnation,[14] I suspect that people have little regard for the world subsequent to their own deaths so that the special case where $D=1$ and $E=0$ particularly merits attention. To simplify matters even further, I set $p=q=1$ (no interperiod mortality).

The intemporal consumption opportunities were derived in Section I and appear in Figure 1 as the dotted c_1 and c_2 lines. The initial point is M which is now interpreted as a Golden Rule point. Will the Rule be violated or will the current generation of parents choose to just replace themselves?

The consumption information in Figure 1 is translated into Figure 2 where it appears as the dotted consumption opportunities line (labelled cc). Starting from the Golden Rule consumption point at M, the current generation of F's can get more c_1 at the expense of c_2 up to the point where $c_1 = \overline{X}_1/(\overline{F}_1+\overline{G}_1)$. Moving the other way, they can get more c_2 at the expense of c_1 up to the point where $w_2=c_2$.

The current generation of parents will choose to just replace themselves, staying on the Golden Rule path, only if $c_1=c_2$

[14] By instant reincarnation, I mean a return to the world as an S just after having been a G.

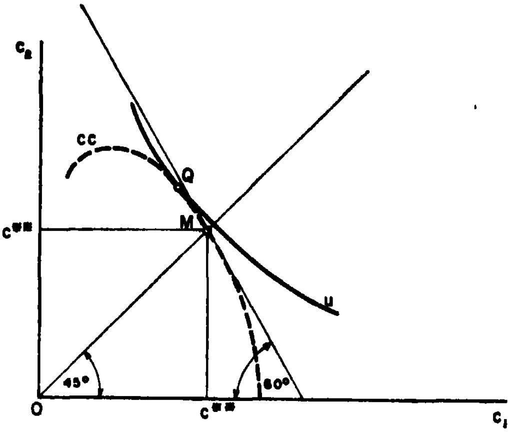


FIGURE 2

$=c^{**}$ maximizes the two-period utility function, equation (1) with $E=0$ and $D=1$. This function can be drawn as a set of intertemporal indifference curves, symmetrical about a 45° ray out of the origin with slopes

$$\frac{dc_2}{dc_1} = -\frac{\mu_1}{\mu_2}$$

At the Golden Rule point, M, $c_1=c_2$ so that $\mu_1=\mu_2$ and

$$\frac{dc_2}{dc_1} = -1$$

The Rule will be violated unless the slope of $cc=1$ at M.

The slope of cc at M is found by dividing (6) by (5), observing that $P_1=P_2$ in a Golden Age and setting $p=1$.

$$\frac{dc_2}{dc_1} = -\frac{w^* - c^*}{c^*}$$

Per capita consumption and per worker product are related by (9). With $p=q=1$

$$c^* = \frac{1}{3}\cdot\left(\frac{X}{F}\right)^*$$

If, in addition, the Golden Rule is being observed, $(X/F)^{**}=w^{**}$ so that $w^{**}=3\cdot c^{**}$ and

$$\frac{dc_2}{dc_1} = \frac{3c^{**} - c^{**}}{c^{**}} = -2$$

is the slope of cc at M.

I conclude that the Golden Rule will be violated. At point M, technology and the share alike ethic permit parents to trade off one unit of consumption in the first period for two in the second. But they are prepared to trade even. Second-period consumption appears relatively cheap to them and they will opt for more of it by procreating in excess of Golden Rule standards, driving c_1 below c^{**} and c_2 above it, moving along cc to point Q. The family moves toward an "overpopulated" condition.

I shall not trace out the adjustment path. That is a very difficult task, even in this simple model.[15] However, the eventual equilibrium must satisfy (10) which, with $p=q=D=1$ and $E=0$, reads

$$w^* = \frac{2}{3}\cdot\left(\frac{X}{F}\right)^*$$

Per capita consumption falls short of its maximum sustainable (Golden Rule) value.

Why does this overpopulation occur? After all, the parents are constrained by the share alike ethic. Maximizing *their own* utility must have a salutory effect on *everyone else's* utility. Or does it? The trouble is simply the fact of human mortality and that people tend to discount the state of the world subsequent to their deaths. In this case, having children imposes costs on the parents when they are parents, generates benefits when the parents have become grandparents, *and* imposes costs once again *after the original parents are dead* and their children have become dependent grandparents. The private fertility decision makers (parents) are able to externalize some of the costs of having children to future generations by dying. The private cost of having children falls short of the social cost, if we can think of the continuing family unit as the "society." This illustrates once again the *fundamental market failure* which exists when the consequences of current decisions carry beyond the decision makers' planning horizon. In this case, the welfare of the children when retired, of their children, and of their children's children are not taken into account.[16]

IV. Private Property and Other Social Contrivances

If the pension motive for having children is strong, one wonders how the desired size of family would change if alternate sources of pensions were available.

Share alike distribution of family product is an obvious method for ensuring that the retired have income during their retirement years. A major rationale for the share alike ethic may be precisely because it leads to automatic pensions in a world where interperiod storage of goods is impossible and financial markets are not developed. But suppose that institutions of private property and market contract spring up to link together contemporary families and to provide intergeneration contracting through marketing of contracts for delivery of goods in the future (pensions). Highly developed institutional forms are not required to perform these functions. A changeover to vesting the ownership of land in *individuals* rather than continuing family units will do the trick. The retired gradually sell off titles to their land to working generations in return for goods. The working generations have an incentive to buy the land, for they

[15] The adjustment path is described by a third (at least) order difference equation. The path is convergent for specifications of the adjustment process which put realistic upper and lower bounds on the control variable S_1. Otherwise, the population is likely not to converge.

[16] Note that this occurs independent of "like father, like son" effects. This model takes the fertility of future generations as exogenous.

look forward to trading their land for goods during their own retirement years. The current purchase of land with goods implies the future delivery of goods for the working generation when they retire.

Money is an alternate social contrivance. Workers buy money from the retired in exchange for goods in anticipation of selling the money for goods while retired. The current purchase of money with goods implies the future delivery of goods for the working generation when they retire.

The impact of these alternate pension schemes on desired family size depends critically upon the extension of the market beyond the family unit. Even if the share alike ethic has withered away, the pension motive for procreation will remain if the "market" for money or land is internal to the family. In that case, parents will perceive the possibility of improving the terms on which they can trade land (or money) for goods to their children, if they have more children (at least up to a point).

The pension motive disappears, however, if markets extend beyond the bounds of the family unit and if the share alike ethic is abandoned. Then having children imposes rearing costs but confers no benefits on the parents. Some other motive would have to account for a continuing population.

Traditional ways are not easily given up, however, and new institutional forms are not likely to be adopted unless it appears advantageous to do so. In particular, the social contrivances I have mentioned, private property and money, will not spring into existence spontaneously. And even though there may be a collective will to reduce population size, and even though these contrivances may serve that end, parents may persist in providing their pensions in traditional ways if their children are good investments compared with available alternatives. How do children compare as investments with land or money?

Consider the simplest case: no intergeneration mortality, no discounting of retirement utilities and no weight given to the future after death. What is the two-period consumption rate of return in equilibrium on land and money?

These contrivances give rise to a transfer of goods from working to retired generations which simulates a costless carry forward of goods. The workers can trade one for one, consumption now for consumption in the future.

But note that the two-period consumption rate of return on children is also zero, under equilibrium conditions in the same model. This is clear from the $c_1 = c_2$ property in any Golden Age and the assumed lack of time preference.

I conclude that the individual fertility decision makers will perceive no clear advantage from abandoning the use of their children for pension purposes, if the alternatives are private property or the use of money, and the share alike ethic will survive.

Perhaps other alternatives would be more attractive. Suppose, for example, that money bore a positive rate of interest. By money, I mean assets issued by agents outside the primitive society: interest bearing liabilities of public or private financial intermediaries. I shall call these liabilities "bonds." An example would be deposit accounts in cooperative credit unions which mobilize saving in a primitive rural sector for investment in a modern industrial sector.

If bonds were available to a share alike society, would parents desire to reduce their family size? The answer is yes—the good asset (bonds) drives out the bad asset (children). An initial Golden Age, before the introduction of bonds, is depicted at point M in Figure 3. The return on children is zero reflecting time preference.

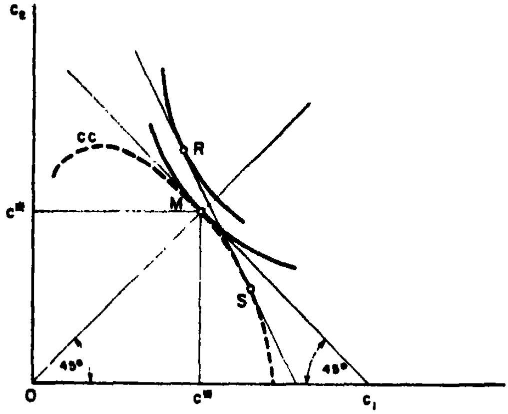


FIGURE 3

Now introduce bonds which bear a positive interest rate. Point M is no longer consistent with utility maximizing, for parents will perceive the possibility of generating investable funds by having fewer children and trading those funds forward at a positive rate of interest. The decline in fertility is denoted by a movement along cc from M to S. The corresponding c_1 and c_2 are consumptions that would be enjoyed if bonds were not purchased. But now the c_1 in excess of c^* can be traded forward by buying bonds and a new consumption point achieved at R, with $c_2 > c_1$ and the return on children equalized with the return on bonds.

This is the reaction of the first generation of parents subsequent to the introduction of bonds: fertility is reduced below Golden Age levels. I infer that subsequent generations will not be satisfied to procreate at levels which would maintain population size. The population will wither away unless there are other motives for having children.

V. Conclusions

Few economists or policy makers would take the view that no important country today is overpopulated in the Golden Rule sense discussed in Section II. Most of us would argue that individual fertility decisions have given too much weight to the quantity of life relative to its quality. I am a professed naive optimalist. I would go all the way to favor the quality of life alone, and opt for the Golden Rule.

My analysis in Section III suggested a fundamental market failure which is rooted in human mortality and a chronic inability of most of us to care much about the state of the world subsequent to our deaths. The effects of the market failure on population size are offset by positive time discounting which assigns a lower weight to utility while old and retired than to utility while young, working, and raising children. The net effect could lead to Golden Rule fertility. But if parents are optimistic of living into retirement, assign relatively high weight to their retirement utility, and care not at all for future generations, the model predicts an equilibrium population in excess of Golden Rule levels on account of the pension motive *alone*.

A more complete model would take uncertainty into account: the uncertainty of survival to retirement and the uncertainty that the children will live on to support their parents in retirement. Risk averting parents would "overprocreate" to reduce the chances of starvation in old age. This effect is attenuated if extended family arrangements or some other social insurance scheme is in effect to take care of the childless retired.

But even the simple certainty model I have worked with goes far enough to demonstrate that one should not rule out the possibility of Golden Age populations in excess of Golden Rule levels being generated by the pension motive *alone*.

This suggests that population size can be significantly reduced if alternate pension schemes are made available. However, they must appear superior to children or they will not be adopted by parents. The analysis in Section IV showed that a posi-

tive rate of return on bonds will eventually eliminate the pension motive (and the model population) if parents give equal weight to their utilities in their two remaining life cycle periods. It is easy to show more generally that bonds must carry a two-period rate of return in excess of the parent's two-period rate of discount in order to have the fertility attenuating effect.

The implications of these observations are two-fold:

They may help explain lower levels of fertility in urban areas where parents have easy access to highly developed financial markets for pension purposes and, perhaps for that reason, family ties are loose and life is impersonal.[17]

If a population control problem is thought to exist, a control scheme might well include the provision of pensions by means of private or social pension plans. A combination of pensions and pills might well be more effective than pills alone.

[17] This implication is hardly new. See, for example, Friedman (p. 208).

REFERENCES

G. Barcley, *Techniques of Population Analysis*, New York 1958.

G. Becker, "An Economic Analysis of Fertility," in *Demographic and Economic Change in Developed Countries*, Nat. Bur. Econ. Res. Conference report, Princeton 1960.

J. Buttrick, "A Note on Growth Theory," *Econ. Develop. Cult. Change*, Oct. 1969, *9*, 75–82.

——, "A Note on Professor Solow's Growth Model," *Quart. J. Econ.*, Nov. 1958, *72*, 633–36.

J. Conlisk, "A Modified Neoclassical Growth Model with Endogenous Technological Change," *Southern Econ. J.*, Oct. 1967, *33*, 199–208.

E. Davis, "A Modified Golden Rule: The Case with Endogenous Labor Supply," *Amer. Econ. Rev.*, Mar. 1969, *59*, 177–81.

M. Friedman, *Price Theory*, Chicago 1962.

H. Leibenstein, *Economic Backwardness and Economic Growth*, New York 1960.

R. C. Merton, "A Golden Golden-Rule for Welfare-Maximization in an Economy with a Varying Population Growth Rate," *Western Econ. Rev.*, Dec. 1969, *7*, 307–18.

R. R. Nelson, "A Theory of the Low-Level Equilibrium Trap in Underdeveloped Countries," *Amer. Econ. Rev.*, Dec. 1956, *46*, 894–908.

J. Niehans, "Economic Growth With Two Endogenous Factors," *Quart. J. Econ.*, Aug. 1963, *77*, 349–71.

E. S. Phelps, "The Golden Rule of Procreation," in his *Golden Rules of Economic Growth*, New York 1966, 176–83.

—— and R. A. Pollack, "On Second Best National Saving and Game-Equilibrium Growth," *Rev. Econ. Stud.*, Apr. 1968, *34*, 185–99.

P. A. Samuelson, "A Catenary Turnpike Theorem Involving Consumption and the Golden Rule," *Amer. Econ. Rev.*, June 1965, *55*, 486–96.

T. P. Shultz, "An Economic Model of Family Planning and Fertility," *J. Polit. Econ.*, Mar./Apr. 1969, *77*, 153–80.

R. M. Solow, "A Contribution to the Theory of Economic Growth," *Quart. J. Econ.*, Nov. 1956, *70*, 537–62.

E. A. Wrigley, "Family Limitation in Pre-Industrial England," *Econ. Hist. Rev.*, Apr. 1966, *19*, 82–109.

A Model of Soviet-Type Economic Planning

By MICHAEL MANOVE*

Each year, planning agencies in the Soviet Union construct an annual economic plan for the calendar year that follows. The annual plan is an important element in the Soviet schema for achieving long-term economic growth. Goals of less detailed long-range plans must be considered in the design of the annual plan. Furthermore, the plan is supposed to expose and remedy weak links in the economic structure. But the most important and obvious purpose of the annual plan is to ensure that the economy will function in a reasonable way from day to day.[1]

The material supply plan is a major component of the Soviet-type annual plan. It specifies aggregate output targets and other production indices from which output targets and certain production indices for individual productive units are derived. The traditional procedure for constructing the material supply plan is very complicated and involves a large planning bureaucracy. In fact, it is difficult to determine from the descriptive literature how rational the planning procedure is, or even how different elements of the procedure fit together.[2] In this paper, an attempt is made to establish a theoretical underpinning for this traditional Soviet-type planning procedure.[3] In particular, we try to analyze how, in theory, a planning procedure of this type can produce a consistent plan, i.e., a plan in which the supply and demand for each commodity is balanced.[4] In order to simplify our task, we shall not consider the Soviet planning system itself. Rather, we shall concentrate our analytical efforts on a more transparent, though not dissimilar, planning system—namely, that of the Autonomous Soviet Republic of Morozhenoe. The planning procedures of this republic oblige us by incorporating only the most basic elements of traditional Soviet-type systems.

I. Characteristics of an Annual Plan

Morozhenoe, incorporated as an Autonomous Republic of the Russian Soviet Federated Socialist Republic shortly after World War II, is the only Soviet republic that has been completely autonomous in the economic sphere. Even more surprisingly, despite the fact that Morozhenoe has a smaller gross product than

* Assistant professor of economics, University of Michigan. I am indebted to Maria Augustinovics, Michael Bruno, Evsey D. Domar, Richard S. Eckaus, Duncan K. Foley, Martin L. Weitzman, and a referee for their extremely helpful comments and suggestions at various stages of this work. I wish to acknowledge financial support and encouragement from the Comparative Economics Program of the University of Michigan.

[1] According to G. Sorokin (p. 225), the long-term plan outlines ways of achieving the main tasks of economic development, the five-year plan elaborates construction and operational plans in greater detail, and the annual plan itemizes the five-year program and facilitates economic management. The annual plan is legally binding.

[2] The most extensive description by a western scholar of Soviet material-supply planning is contained in the unpublished doctoral dissertation of Herbert Levine (1961). A condensation of the dissertation appears in his 1959 paper. I. A. Evenko (pp. 72–88) is another source of general descriptive material. For a description of the Polish procedure, see J. M. Montias (1962, pp. 6–32, 76–114).

[3] An excellent survey of possible theoretical approaches is contained in Montias (1959).

[4] For general information on the consistency of Soviet plans, see Michael Ellman. Montias (Oct. 1962) discusses the possibility of utilizing the decomposability of the A-matrix to produce consistent plans.

even Darien, Connecticut, this little republic is virtually self-sufficient and engages in little external trade.

Wishing to embark on a program of rapid industrialization, the Morozhenoens decided to construct a "command economy," of which an annual plan of material supply was to be an integral part. Because the system envisioned was to be basically non-market, the annual plan would have to serve as the main guide for the producing units of the economy in their day-to-day operations. The system for constructing such short-term material supply plans, it was decided, ought to have the following characteristics:

1) The authorities specify net output levels of final goods (including capital goods and planned inventory accumulations).

2) The system yields a plan which determines, at the very least, a set of output targets for each producing unit.

3) The system yields a plan which is internally consistent. For each output target specified by the plan, sufficient quantities of the necessary inputs must be provided.[5]

4) The system yields a plan which is reasonably efficient in the short run. To the extent permitted by long-term constraints, existing resources and factors of production should be put to full and reasonable use. If the necessary labor is available, a short-term plan ought to set output targets at levels which utilize the full capacities of most producing units. The Morozhenoens did realize, however, that the scope of an annual plan is too limited to deal with questions of efficiency involving the allocation of capital or major shifts in technology.

II. The Input-Output Planning Procedure

The Morozhenoens have had a long standing aversion to Russian bureaucratic planning methods. In 1947, risking Moscow's ire, they invited a well-known American economist from a major New England academic institution to design for them a simple, non-bureaucratic material supply planning system. They stipulated that the system have the previously discussed characteristics. The American advised the use of the following input-output procedure.

Let y_j be the gross output target for commodity j, d_j the final demand for commodity j, and a_{ij} the input of commodity i required per unit output of commodity j (a fixed-proportions production function is assumed), and let Y, D, and A denote the corresponding vectors and matrix, respectively. Suppose that A and D are known. To determine a consistent set of gross output targets, gross supplies are equated with gross demand:[6]

$$Y = AY + D \tag{1}$$

Solving for Y, we obtain

$$Y = (I - A)^{-1}D \tag{2}$$

where I is the identity matrix. The vector Y may also be given inductively, as in the following:

$$Y_{(i)} = AY_{(i-1)} + D \tag{3}$$

where $Y_{(i)}$ approaches Y as i gets large.

Equation (3) specifies what we shall call a "supply-demand iteration"; it sets a new output (supply) vector equal to the demand vector associated with a previous output vector. To use (3), one starts out by specifying a tentative gross output vector, say $Y_{(0)}$. The right-hand side of (3) yields the demand vector associated with

[5] This is only one sense in which a plan may be said to be consistent. Richard Stone provides a fairly complete summary of the concept of consistency in planning.

[6] In this paper, the terms "output" and "supply" are used interchangeably. Imports and planned decreases in inventory stocks are considered to be negative summands of final demand. The demand for an intermediate good is completely determined by the output targets and the fixed-proportions production functions and is assumed to be strictly independent of price.

the production of $Y_{(0)}$; namely, $AY_{(0)}+D$. A new tentative output vector $Y_{(1)}$ would then be defined as equal to this demand vector. Equation (3) would be applied repeatedly, yielding $Y_{(2)}$, $Y_{(3)}$, and so on.

If any given output vector $Y_{(i)}$ were produced, then the corresponding demand vector would be $AY_{(i)}+D$. Let $E_{(i)}$ be defined as the supply-demand imbalance associated with the production of $Y_{(i)}$, that is, the difference between the supply $Y_{(i)}$, and the demand $AY_{(i)}+D$. Then, the imbalance associated with $Y_{(i)}$ is given by

$$E_{(i)} = A^i E_{(0)}$$

Since A is productive,[7] we know that A^i approaches 0 as the exponent i gets large. It follows that with succeeding supply-demand iterations $E_{(i)}$ also approaches 0 and that $Y_{(i)}$ goes to Y. (How fast does the supply-demand imbalance approach 0? A trial with a 38-sector A-matrix for the USSR and with $E_{(0)}$ proportional to gross output, revealed that in general, each iteration cut the supply-demand imbalance roughly in half.)[8]

If the demand for final products, D, and the input norms, A, are known to a central planning agency, the agency could grind out Y with a computer in a relatively short period of time using either (2) or (3). It would remain to disaggregate the gross output targets with respect to individual enterprises. In Morozhenoe, however, because of the relatively small number of enterprises, this would present few difficulties.

The input-output procedure is extremely attractive because of its simplicity and because plans produced by it are perfectly consistent. And while the above planning method does not imply short-run efficiency (which depends on the values of A and D), it is not incompatible with it. Unfortunately, the input-output procedure could not be successfully implemented in Morozhenoe.

The principal difficulty with the scheme was that all information on A and D had to be known by the center in order to be used. The relevant information had to be obtained in great detail at the local level and then summed up for the entire economy, a task which overwhelmed Morozhenoen information processing facilities. Even for this small 1000-good economy, there are a total of 1,001,000 entries in A and D, of which roughly 200,000 are non-zero.[9] And since the amount of data to be

[7] A technology matrix A is said to be productive if a positive vector of net outputs is producible; i.e., A is productive if, and only if, there exists Y such that $Y-AY>0$ (in every component). It can be shown that A is productive if, and only if, $(I-A)^{-1}$ exists and is nonnegative.

[8] A 38-sector A-matrix and a vector Y of gross outputs for the Soviet Union were obtained from V. Treml (1966). Starting with supply-demand imbalances proportional sector-by-sector to gross outputs, successive supply-demand iterations (3)—or, equivalently, successive multiplications of the imbalance vector by the A-matrix—reduced the ruble value of the supply-demand imbalances by an average of 51.0 percent per iteration. In the sector with the smallest average reduction, the imbalance was reduced by 28.4 percent per iteration. Had we assumed that the imbalances in different sectors had differing signs, the reductions would have been considerably greater. (Note: successive iterations do not reduce the imbalance in a given sector by a constant percentage; the above figures are arithmetic averages, with each percentage reduction weighted by the size of the imbalance before the iteration.)

It is unlikely that the degree of aggregation of the A-matrix and the imbalance vector would have a significant effect on the rate at which multiplication of the imbalance vector by the A-matrix reduces that vector. The above experiment was repeated with a 17-sector version of Treml's matrix, and the results were virtually unchanged. Indeed, if we assume that the A-matrix was derived from an input-output table, and if the original imbalance vector is proportional to the gross output vector given in that table, then it can be shown that the percentage change in total value of the imbalances as a result of multiplication by the A-matrix is independent of the degree of aggregation.

[9] A crude estimate. In Eidel'man (p. 237), it is reported that 4,754 non-zero entries were contained in 83-sector Soviet 1959 A-matrix. Assuming that the number of non-zero entries in an A-matrix is greater than proportional to the dimension of the matrix and less than proportional to the number of entries in the

compiled grows more than proportionally to the number of commodities produced, the difficulty of this task increased rapidly as the economy grew and became more complex.[10]

By 1953, trial use of the input-output planning procedure was discontinued. The stock of unwanted inventories equaled almost two-thirds the size of the annual gross national product, and shortages abounded. The leading Morozhenoen economists realized that this situation resulted not from any error internal to the planning procedure, but rather from the use of incorrect values in the A-matrix and in the final demand vector D. Some economists argued that the techniques of information gathering should be improved and that the input-output planning system be retained. But they were overruled by the authorities who insisted that the planning system be scrapped in favor of a new system that requires much less information in order to function. The Morozhenoen government set up an emergency committee to draw up an annual planning procedure based on Russian practices at the time. On October 20, 1953, this new procedure was formally adopted.

III. The Soviet-Type Planning Procedure in Morozhenoe

The Soviet-type planning procedure depends on the existence of a large planning bureaucracy, both at the central and at the enterprise levels. All commodities are placed in either of two categories: centrally planned commodities and locally planned commodities. Information on locally planned commodities is processed entirely at the enterprise level, while information pertaining to centrally planned commodities is passed back and forth between the central planning agency and the enterprises.

In Soviet Morozhenoe, the central planning agency is known as Gosplan. Gosplan has a series of departments organized along industrial lines. Each centrally planned commodity is assigned to two of these departments: a so-called summary department, and an industrial department. The summary departments are responsible for processing information dealing with demand and distribution of commodities, while industrial departments deal with the supply of commodities from production and other sources. Gosplan also contains statistical departments and final-demand departments. How the various departments of Gosplan interact with each other and with local planning agencies is described on the following pages.

Conceptually speaking, the Soviet-type planning procedure may be thought of as a series of iterations. These iterations are of three types, all of them variants of the basic supply-demand iteration described by equation (3). One of these, the "retrospective iteration," may be used by local planning agencies in determining output targets for locally planned commodities, and by central planning agencies in determining output targets for centrally planned commodities. The other two, the "external iterations" and the "internal iterations," require certain centrally processed information, and, consequently, may be used to adjust the output targets of only centrally planned commodities. In Morozhenoe, the process of drawing up an annual plan spans a one-year period.

The principal steps of the annual planning procedure are outlined below (see, also, Figure 1) and a mathematical representation for the procedure is presented.

Let n denote the number of commodities

---

matrix, a guess of 200,000 as the number of non-zero elements in the 1000x1000 Morozhenoen A-matrix was derived by taking a geometric average of 4754x1000/83 and $4754 \times (1000/83)^2$.

[10] The difficulties of applying input-output methods to Soviet planning are outlined by A. Dorovskikh (pp. 38–39).

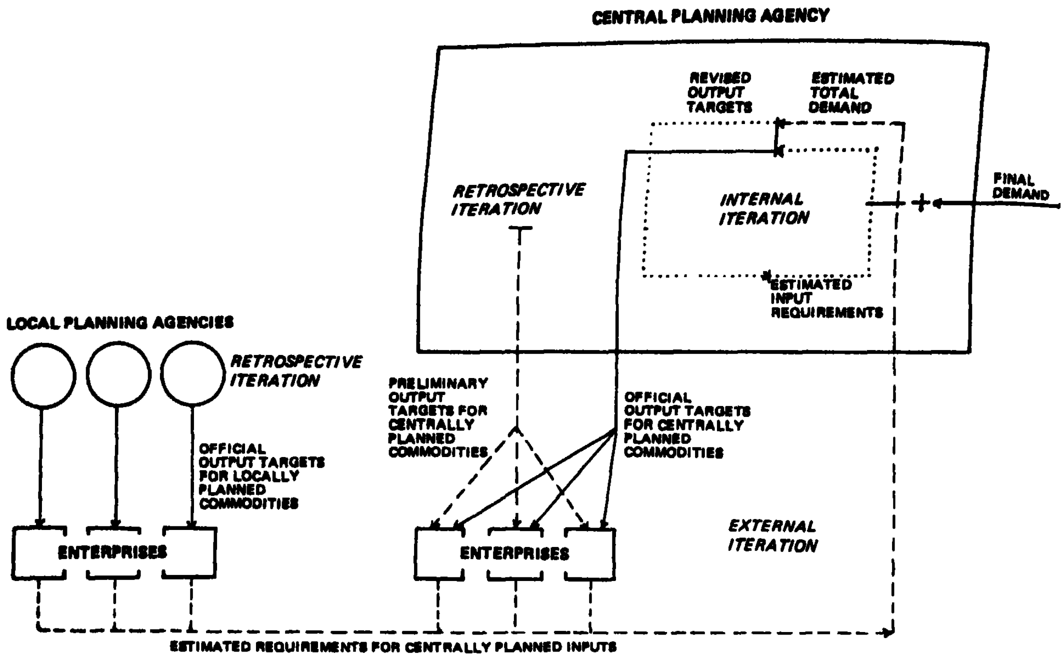


FIGURE 1—SOVIET-TYPE MATERIAL-SUPPLY PLANNING PROCEDURE: A STYLIZED VERSION

being produced. All of the vectors defined below are n-dimensional; the ith component of a vector always pertains to commodity i. If the vector is subscripted, the subscript refers to the year for which the vector is defined. Each commodity is produced by many enterprises. The word "aggregate" is used to indicate a nationwide total pertaining to a given commodity, and *not* the combining of different commodities. A list of symbols used appears in the Appendix.

Step 1: Specification of preliminary aggregate output targets: a retrospective iteration (January–March).[11]

For each centrally planned commodity, the center sets a preliminary aggregate output target by adding an exogenously determined increment in gross output to the previous year's aggregate demand.[12] The preliminary aggregate output targets (control figures) are then used to assign output targets to individual enterprises.

No preliminary aggregate output targets are set for locally planned commodities. Instead, a preliminary output target is set independently for each enterprise producing a given commodity, by adding an exogenously determined increment in production to the demand for the commodity on the enterprise (i.e., the orders placed with the enterprise) during the previous year. The reader should note that while a preliminary output target is not set, or known, for a locally planned product, such a target does in fact exist—namely, the sum of all of the enterprise output targets for the product.

For both centrally and locally planned

[11] The time periods given here are similar to those for the Soviet Union as reported in Evenko, (p. 84).

[12] It is assumed that in cases where actual requirements for centrally planned inputs differ from planned allotments, producers will ask the center to make appropriate adjustments in the allotments. These requests for adjustments enable the center to calculate actual aggregate demand for centrally planned products after the fact.

commodities, preliminary output targets (potential supplies) are based on actual demands of the previous year. Thus the designation retrospective iteration.

Let $\hat{Y}_t$ be the vector of preliminary aggregate output targets for year t. Then, in accordance with the above specifications, $\hat{Y}_t$ is given by the vector equation

$$\hat{Y}_t = X_{t-1} + Q_t$$

where Q_t is a vector of exogenously determined increments (or decrements, if negative) to aggregate outputs,[13] and X_t is the vector of aggregate demands. For analytical purposes, however, we shall allow the possibility of omitting the retrospective iteration for some commodities, in which case the previous year's output would be substituted in the above equation for the previous year's demand. Therefore we amend the equation to

$$\hat{Y}_t = \Omega X_{t-1} + (I - \Omega) Y_{t-1} + Q_t \quad (4)$$

where Y_t is the vector of aggregate outputs, Ω is a diagonal matrix whose ith entry is equal to 1 if commodity i is to be iterated retrospectively, and 0 otherwise. All quantities indicated in (4) are aggregate quantities.

Equation (4) is valid for both centrally and locally planned commodities, but for a different reason in each of the two cases. It is true for centrally planned commodities because the right-hand side of the equation actually describes the way the center sets preliminary aggregate output targets. Equation (4) is valid for locally planned commodities, because an equation analogous to (4) but in terms of non-aggregate quantities, defines each enterprise target. Summing all of the non-aggregate equations pertaining to a particular locally planned commodity yields the relationship for aggregate quantities expressed by (4).

Step 2: Computation of tentative demands (April–June). The vector $\hat{X}_t$ of tentative aggregate demands induced by the preliminary output targets $\hat{Y}_t$, is given by

$$\hat{X}_t = A\hat{Y}_t + D_t \quad (5)$$

where D_t is the vector of planned final demands. Gosplan, however, cannot use this equation to compute the tentative demands, since it does not necessarily know the value of the A-matrix or the values of $\hat{Y}_t$ and D_t pertaining to locally planned commodities. Instead, Gosplan calculates aggregate demands for centrally planned products only. The following procedure is used:

All enterprises submit estimates of the quantities of all *centrally planned* commodities they plan to use as inputs.[14] In making these estimates, each enterprise assumes its output will equal its preliminary output target. The estimates of required inputs are aggregated and forwarded to Gosplan. At the same time, final demands for centrally planned products are set by final demand departments of Gosplan, which, at its discretion, may be guided by consumers' and investors' desires.[15] Each summary department of Gosplan then determines estimates of aggregate demands for the centrally planned products assigned to it by adding the demand for each product as an input to the respective final demand.

Step 3: Revision of the preliminary output targets for centrally planned commodities: an external iteration (July).

At this point, the summary departments of Gosplan consult with the corresponding industrial departments. If the tentative

[13] The centrally planned components of Q_t are set explicitly by Gosplan; the locally planned components are the sums of exogenous changes made by local planning agencies.

[14] For a description of this process in the USSR, see Levine (1959, pp. 156–59).

[15] The planning of consumption is treated at length by Phillip "'eitzman.

demands do not equal the tentative supplies—and they surely will not—then the demand and supply figures must be reconciled. *In order to effect this reconciliation, Gosplan uses the material balance, an accounting device in which all supplies of a commodity, on the one hand, and all demands for the commodity, on the other, are tabulated. In most Socialist countries*, the process of adjusting supplies and demands in order to bring about a balance is extremely complicated and draws heavily on the experience and intuition of the planners. In Soviet Morozhenoe, however, only one basic aspect of the process is used. The preliminary output targets for centrally planned commodities are reset to equal the respective aggregate demand estimates. Thus a new tentative aggregate gross output target is formulated for each centrally planned commodity.

The output targets for locally planned commodities remain unchanged.

Let $Y_{t(0)}$ be the revised vector of aggregate gross output targets. It follows from the above that $Y_{t(0)}$ will equal X_t in centrally planned components and $\hat{Y}_t$ in locally planned components. These relationships can be expressed as one equation:

$$Y_{t(0)} = \Gamma \hat{X}_t + (I - \Gamma)\hat{Y}_t \tag{6}$$

where Γ is a diagonal matrix with entries pertaining to centrally planned commodities set equal to 1, and entries pertaining to locally planned commodities set equal to 0.

As far as centrally planned commodities are concerned, Steps 2 and 3 have the same result as one supply-demand iteration (3) of the input-output procedure, in that the demands generated by the preliminary output targets become new output targets. With the input-output procedure, however, the center must calculate the demands for inputs from its knowledge of production functions (the A-matrix), while in Step 2 of the Soviet-type procedure, the center (Gosplan) allows the producing units (enterprises) to estimate their own input demands. Because part of the demands are estimated outside of the center, Step 3 is referred to as an external iteration.

Step 4: Further revisions of tentative output targets for centrally planned commodities:[16] internal iterations (August).

In principle, the internal iteration is a straightforward process. Whenever the tentative aggregate output targets for centrally planned commodities are revised, the tentative input requirements of the producers of centrally planned commodities will change. When an industrial department of Gosplan changes (or is advised of a change in) output targets of commodities under its purview, it may advise appropriate summary departments of resulting changes in the input requirements of the industry, provided, of course, that the information necessary to compute the changes in requirements is available. After the summary departments receive notification of changes in input requirements from all of the industrial departments, the summary departments may proceed to calculate a revised set of aggregate demands for the centrally planned commodities. The summary departments may then consult with the corresponding industrial departments, so that the latter can set new aggregate output targets equal to these revised aggregate demands.

Let $Y_{t(i)}$ denote the tentative aggregate output targets after i internal iterations. Then we may represent an internal iteration by

$$Y_{t(i)} = \Gamma X_{t(i-1)} + (I - \Gamma)\hat{Y}_t \tag{7}$$

where $X_{t(i)}$, the vector of tentative ag-

[16] That a realistic model of Soviet-type planning ought to include this step was suggested by Martin Weitzman, Yale University, and by Maria Augustinovics, Head, Division of Long-term Planning, National Planning Institute, Budapest, Hungary.

gregate demands after i internal iterations, is given by

$$X_{t(i)} = A Y_{t(i)} + D_t \tag{8}$$

The internal iteration also yields results like those of the supply-demand iteration (3), except that in this case, only the output targets of centrally planned commodities are revised. The name internal iteration derives from the fact that all computation is performed within Gosplan, itself. In order for the internal iteration to be carried out, the industrial departments must have certain information about input usage in the production of commodities under their purview. In particular, they must know the specific coefficients in the A-matrix which yield the quantities of centrally planned inputs needed to produce these commodities.

Step 4 consists of a number of internal iterations, a number bounded by time, cost and information constraints on Gosplan. Because of these constraints, some of these internal iterations may have to be partial,[17] or all internal iterations may even be omitted.

At this point, the reader may be wondering why an external iteration is necessary given that subsequent iterations may be internal. The explanation lies in the fact that the first calculation of tentative aggregate demands for centrally planned commodities requires more information than subsequent calculations do. The first calculation requires information on centrally planned inputs used to produce *all* commodities, both centrally and locally planned. But, because the tentative output targets for centrally planned commodities are the only ones that change in the iterations, subsequent calculations of the tentative aggregate demands require information pertaining only to the use of centrally planned inputs in the production of centrally planned commodities. Furthermore, the center may be able to obtain much of the information needed to execute internal iterations as a by-product of the external iteration.

Step 5: Specification of the official output targets (September–December).

The most recently derived tentative aggregate gross output targets for centrally planned commodities become official targets after approval by various government bodies. These targets are then disaggregated and assigned to the individual enterprises producing centrally planned commodities. Official output targets for locally planned goods are the same as the preliminary targets set in Step 1.

If n internal iterations were carried out, the official output targets $\bar{Y}_t$ are given by

$$\bar{Y}_t = Y_{t(n)} \tag{9}$$

It is clear that the Soviet-type planning procedure is less demanding with regard to information gathering than is the input-output procedure. The input-output procedure required the center to know all of the elements of the A-matrix, to calculate total output and demand for each commodity, and to know or set total final demand for each commodity. The Soviet-type procedure, however, requires only that the center calculates total output and demand for centrally planned commodities and that it knows or sets final demand for centrally planned commodities. If internal iterations are omitted, knowledge of the A-matrix by the center is not required. However, if the center does know coefficients pertaining to the use of centrally planned inputs in the production of centrally planned commodities, it may use them to

[17] For the sake of mathematical simplicity in analyzing the model, I have assumed that an internal iteration involves all centrally planned commodities. A more general model would allow internal iterations to be performed for any subset of the centrally planned commodities. This would permit the center to carry out partial internal iterations which would require the knowledge of technical coefficients only for those commodities involved.

carry out internal iterations, thereby increasing the degree of consistency of the plan—as we shall see.

As the Soviet-type procedure was being established in Morozhenoe, many economists there predicted that plans produced by the procedure would be subject to catastrophic failure. It was agreed that the system does allow (but does not require) the authorities to set the level of final demand for a commodity, and does yield output targets for each producing unit. There was some concern about short-run economic efficiency. But the major cause of skepticism centered around the question of consistency.[18] Why, it was asked, should the planned supply of a commodity equal the resulting demand for it? For one thing, only about one-third of all commodities were to be centrally planned. Considering the fact that current output targets for locally planned goods are not directly coordinated with the current output targets for centrally planned goods, what assurances are there that the supplies of locally planned goods required as inputs throughout the economy would be available? The usefulness of the external and internal iterations seemed doubtful for two reasons: the planners would have time and funds available for at most a few iterations, and only centrally planned output targets were being revised. It was suggested that the abolition of all central planning would be preferable to these halfway measures.

Between 1954 and 1956, the chaotic state of the Morozhenoen economy seemed to bear out the pessimism of the economists. But, surprisingly, the plan of 1957 turned out to be almost consistent. The supply and demand for each centrally planned product differed at most by a few percentage points, and the supply and demand for each locally planned product differed by no more than about 5 percent. And this, despite the fact that internal iterations had been entirely omitted from the planning procedure for economy reasons. Results in later years were just as good. An analysis of the model suggests an explanation.

IV. An Analysis of the Soviet-Type Procedure

It is assumed that the relationship between actual outputs Y_t, demands for the outputs X_t, and final demands D_t, is given by

$$X_t = AY_t + D_t \tag{10}$$

The vector of supply-demand imbalances E_t is given by

$$E_t = Y_t - X_t \tag{11}$$

For the purposes of this analysis, we make the simplifying assumption that the planned output targets are always exactly fulfilled, i.e.,

$$Y_t = \bar{Y}_t \tag{12}$$

This implies that E_t will be generated only by the inconsistencies in the plan.

The assumption that output targets are precisely met also implies that the supply-demand imbalances represented by E_t manifest themselves as economic occurrences with little or no secondary effects: perhaps as unplanned changes in the levels of consumption,[19] in the levels of exports or imports, or, under certain circumstances, in the size of inventory stocks. In

[18] Problems of feasibility are avoided by assuming that the Morozhenoens always choose final demands so as to fall within the production possibility curve. Thus if a plan is internally consistent, it is also feasible. The problem of feasibility is treated more realistically in the author's doctoral dissertation.

[19] Of course, unplanned (or any other) changes in the level of consumption will have important secondary effects outside the sphere of production. And if the changes are large enough, or if the population is close to the subsistence level, changes in consumption may effect labor, and consequently, production.

other words, the entire error falls on final demand. Although these assumptions are unrealistic (even in Morozhenoe), they serve to isolate the direct effects of inconsistent plans from the secondary effects, which depend to a large extent on how the economic administration handles the shortages or surpluses that arise from the inconsistencies.

The vector E_t tells the complete story as far as consistency of the plan is concerned. If $E_t=0$, then the plan is perfectly consistent, if the components of E_t are large (+ or −), then large surpluses and/or shortages are indicated. We attempt to throw some light on the genesis of E_t in the Morozhenoen planning procedure by solving the mathematical representation of that procedure. For simplicity, we shall assume that no internal iterations (Step 4) are carried out. Under this circumstance, equations (6)–(9) and (12) reduce to

$$(13)\qquad Y_t = \Gamma\hat{X}_t + (I - \Gamma)\hat{Y}_t$$

The model of the planning procedure with internal iterations is analyzed in the Appendix.

Equations (4), (5), (10), (11), and (13) can be solved simultaneously for the supply-demand imbalance E_T realized in a given year T. If, as we may do without loss of generality, we set $Y_0=D_0=0$, the solution is as follows:

$$(14)\qquad E_T = \sum_{t=1}^{T} (\Gamma_{[1]}\Omega_{[1]})^{T-t}\Gamma_{[1]}L_t$$

where

$$(15)\qquad L_t = Q_t - (AQ_t + \Delta D_t)$$

$$(16)\qquad \Gamma_{[1]} = I - (I - A)\Gamma$$

and

$$(17)\qquad \Omega_{[1]} = I - (I - A)\Omega$$

and where $\Delta D_t=D_t-D_{t-1}$.

Equation (14) has a helpful heuristic interpretation. Note that (14) gives the supply-demand imbalance E_T for year T as a weighted sum of L_t's. What does L_t represent? Recalling that Q_t is the exogenously determined increment in gross output for year t (see equation (4)), we see that $AQ_t+\Delta D_t$ is the exogenously induced increment in aggregate demand. It follows from (9), then, that L_t is the magnitude of the supply-demand imbalance newly generated in year t by disproportions between the exogenously induced increment in gross output, on the one hand, and the exogenously induced change in demand, on the other. Thus, equation (14) breaks down the supply-demand imbalance E_T into a sum, and each term $(\Gamma_{[1]}\Omega_{[1]})^{T-t}\Gamma_{[1]}L_t$ of the sum may be thought of as that portion of the imbalance that was generated in year t.

It turns out that any reasonable specification of the parameters Γ and Ω will guarantee that $(\Gamma_{[1]}\Omega_{[1]})^{T-t}\Gamma_{[1]}$ approaches 0 geometrically, as T−t increases. (Setting $\Omega=I$, its normal value, is sufficient to assure this.) It follows that changes made long ago do not contribute significantly to the current imbalance. Given a uniform degree of consistency of the exogenous increments to output and demand, the degree of consistency of the annual plans will remain at a kind of stable equilibrium. There is no positive feedback endogenous to the system which could result in a deterioration of the consistency of the plans from year to year. It is possible, of course, that the consistency of the exogenous increments to output and demand could steadily worsen over time, i.e., the differences between the increments to output and increments to demand could become a larger and larger percentage of the size of the increments to output. But even relatively large imbalances in the exogenous increments to output and demand, do not generate major inconsistencies in the annual plans (see examples in Section VI).

The size of the imbalance in the exogenous increments to output and demand is not the only determinant of the degree of consistency in the plans. The degree of consistency also depends on how many and which goods are centrally planned and on the number of internal iterations in the planning procedure, as we show below.

V. Some Specific Solutions for Plan Inconsistencies

What does (14) look like for various values of the parameters Ω and Γ? If a retrospective iteration is performed for all commodities ($\Omega=I$), then with no central planning at all ($\Gamma=0$), we have

$$E_T = L_T + AL_{T-1} + A^2L_{T-2} + \dots + A^{T-1}L_1 \tag{18}$$

In this case, the portion of the supply-demand imbalance of year T that was generated in year t is $A^{T-t}L_t$. Thus, since A is productive, the longer the elapsed time from the introduction of exogenous increments in gross output and final demand, the smaller, in general, will be the supply-demand imbalances attributable to those changes. (For an actual A-matrix, multiplication of a vector by A^i reduced most components of that vector by a factor of about 2^i (see footnote 8). E.g., the part of a supply-demand imbalance generated by output and demand increments made five years previously would be less than roughly 3 percent of the imbalance that would have been caused by the same increments at the time they were introduced.)

This solution bears an interesting relationship to the supply-demand iteration procedure (3). Suppose that supply-demand iterations were used to balance the exogenously determined increment in supply Q_t with an exogenous increment in final demand ΔD_t. From (3) we get

$$Q_{t,i} = AQ_{t,i-1} + \Delta D_t$$

where $Q_{t,i}$ is defined to be Q_t as revised by i iterations (with $Q_{t,0}=Q_t$). Solving this as a difference equation yields

$$Q_{t,i}=A^iQ_t+(I+A+\dots+A^{i-1})\Delta D_t \tag{19}$$

Now, if, for a given i, the output increment $Q_{t,i}$ alone were produced, the supply-demand imbalance resulting from that production would be given by

$$Q_{t,i} - (AQ_{t,i} + \Delta D_t)$$

which, by (19), equals

$$A^i[Q_t - (AQ_t + \Delta D_t)]$$

and, by (15), this reduces to A^iL_t. But this term has the same structure as the terms on the right-hand side of (18). Thus, the Soviet-type planning procedure under these assumptions works as if it treats the supply and demand of a given year, not as a whole but as a series of annual exogenous increments which have accumulated over time. And it is as if the planning procedure uses supply-demand iterations to balance each set of annual increments separately from the other sets, with the number of iterations carried out in each case equal to the number of years having passed since that set of increments were introduced.

Suppose, now, that all commodities are centrally planned ($\Gamma=I$), but that the retrospective iteration is completely omitted ($\Omega=0$). Then

$$E_T = AL_T + A^2L_{T-1} + \dots + A^TL_1 \tag{20}$$

This procedure works much the same way as the procedure associated with (18), except that the balancing process for each set of increments includes one more supply-demand iteration than is the case in (18). And this is not surprising. Under the assumptions pertaining to (18), the whole planning procedure amounts to a single retrospective iteration for all commodities. Under the assumption of central planning for all commodities, which yields (20), the

procedure amounts to a single external iteration for all commodities. And the only difference between the retrospective iteration and the external iteration is that the latter takes account of the *currently planned* increments in output and final demand while the former does not.

When the planning procedure specifies central planning for all commodities (i.e., an external iteration) on top of a complete retrospective iteration ($\Gamma=I$, $\Omega=I$), the resulting supply-demand imbalance for year T will be given by

$$E_T = AL_T + A^3L_{T-1} + A^5L_{T-2} + \ldots + A^{(2T-1)}L_1 \quad (21)$$

It is as if each annual set of exogenous increments were balanced with two supply-demand iterations per year, instead of one. The incorporation of central planning for all commodities into the planning procedure more than halves the number of years it takes a newly generated supply-demand imbalance L_t to damp out.

When the planning procedure specifies central planning for all commodities and includes internal iterations, it is as if each annual set of exogenous increments were subject to even more supply-demand iterations per year (see the Appendix). If the planning procedure included three internal iterations, for example, the supply-demand imbalance E_T would be given by

$$E_T = A^4L_T + A^9L_{T-1} + A^{14}L_{T-2} + \ldots + A^{(5T-1)}L_1 \quad (22)$$

While the above cases illustrate the properties of the retrospective and the external and internal iterations, a more realistic example of the Soviet-type procedure is provided by assuming central planning only for some commodities on top of a complete retrospective iteration in setting the initial tentative targets. In this case, $\Omega=I$, but Γ would contain both 1's and 0's on its diagonal. For simplicity we again assume that no internal iterations are performed. With these assumptions, (14) yields the following solution for E_T:

$$E_T = \bar{A}L_T + (A\bar{A})\bar{A}L_{T-1} + (\bar{A}A)^2\bar{A}L_{T-2} + \ldots + (\bar{A}A)^{(T-1)}\bar{A}L_1 \quad (23)$$

where $\bar{A}$ is constructed by taking columns corresponding to centrally planned commodities from the A-matrix, and columns corresponding to locally planned commodities from the identity matrix.

Equation (23) implies the obvious fact that each annual set of increments in output and demand for centrally planned commodities is subject to one more supply-demand iteration than those of locally planned commodities, and it follows that even without using internal iterations, the supply-demand imbalance of a centrally planned commodity would be considerably smaller than that of the commodity were it locally planned. Equation (23) also implies the less obvious fact that the central planning of some commodities generally reduces the supply-demand imbalances of the other commodities which remain locally planned.

VI. Some Illustrative Numerical Solutions for Plan Inconsistencies

Having explored the properties of solutions for the supply-demand imbalance under several specifications of the planning procedure, we try to get more of a hold on the size of the indicated supply-demand imbalances by using real data and a few assumed numbers. In what follows, we assumed that the annual exogenous increments in gross output Q_t are on the order of 10 percent of the total gross output, and that the supply-demand imbalances inherent in these changes L_t are about 25 percent of the size of the increments, and we used Treml's 38-sector version of the 1959 Soviet A-matrix to evaluate our formulas for E_T. In evaluating these for-

mulas, T was chosen sufficiently large to make E_T converge to its steady-state value.

For the case where the planning procedure incorporates only a retrospective iteration, we learned, by evaluating (18), that the supply-demand imbalances average 4.9 percent of the size of the corresponding gross outputs. The sector with the largest imbalance had an imbalance of 8.8 percent.

When all commodities are retrospectively iterated and centrally planned (with no internal iterations), the imbalances average about 1.6 percent of the size of the gross outputs (from (21)); and the addition to the planning procedure of three internal iterations lowers the size of the imbalances to an average of .14 percent of the gross outputs (from (22)).

With half of the sectors centrally planned (no internal iterations used), the supply-demand imbalances of the centrally planned sectors average 1.7 percent of the size of the respective gross outputs as compared to an average of 4.5 percent for the same sectors with no central planning. Also, centrally planning half of the sectors lowers the imbalances in the remaining sectors (still locally planned) from an average of 5.4 percent to an average of 4.9 percent.[20]

VII. Conclusion

In order to gain some insight into the workings of the Soviet-type system of material supply planning, I have constructed a model of "Morozhenoen" planning, which is, of course, a stylized abstraction of the Soviet system, itself. The model was designed with several well-known descriptions of the Soviet procedure in mind, including those of Levine (1959, 1961), Montias (1959), and Evenko. The input-output procedure, which was explicated for the purpose of comparison with the Soviet-type system, is standard.

The input-output method of setting output targets is very attractive because in either its matrix-inversion form (2), or its iterative form (3), virtually perfect balance between supply and demand can be achieved, at least in theory. Unfortunately, in order to use the input-output procedure for detailed material supply planning, the center must have available to it a large amount of information which is difficult to obtain and process. This fact renders the input-output planning procedure of doubtful value to economies which lack advanced facilities for data collection and processing.

The traditional Soviet-type material supply planning procedure, on the other hand, takes advantage of detailed information known on the local level but not known by the central planning agency. Information concerning only "locally planned" commodities need not be known to anyone outside of the producing enterprise. And even with regard to centrally planned commodities, much less information is needed by the center with the Soviet-type procedure than would be needed with an input-output procedure. Yet, as I hope the model has shown, the Soviet-type procedure can, in principle, yield a reasonably consistent plan of material supply. Moreover, as Montias (1959, pp. 967–68) argued years ago, the underlying workings of the system are similar to the iterative version of the input-output procedure defined by (3).

The abstract Soviet procedure as I defined it, was shown to be operationally equivalent to a succession of three kinds of iterations. In the first of these, the retrospective iteration, preliminary output targets are set on the basis of actual demands of the preceding year. This iteration may be carried out entirely on the local

[20] In this example, the largest percentage imbalance among the nineteen centrally planned sectors is 5.7 percent, and the largest among the nineteen locally planned sectors is 6.9 percent.

level for locally planned commodities and entirely on the central level for centrally planned commodities. In the external iteration which follows, revised output targets for centrally planned commodities are set on the basis of demand estimates made in part on the local level assuming the original preliminary output targets. Finally, there may be some internal iterations in which revised output targets for centrally planned commodities are calculated within the planning center itself.

But how can only a few iterations yield a reasonably consistent plan? The answer is that the effect of the iterations accumulates from year to year, a phenomenon brought about by the fact that each year's planning procedure uses previous plans and production and demand figures as a point of departure. As is reflected in the solutions for the supply-demand imbalance derived in the previous section, it is only the recent changes in output and demand that are iterated only a few times. The bulk of these magnitudes were iterated over and over again in previous years. This point, I think, was missed by Levine when he stated the following:

> It is frequently thought that the iterative approach is the basic method used by Soviet planners to achieve consistency in their plans. I do not think that this theory is correct. . . . On the basis of *very crude* calculations it might be said that somewhere between 6 and 13 iterations would be required. It is inconceivable that Gosplan, under the conditions which prevailed could have performed that number of iterations.[21]

Of course, as Levine asserts elsewhere in his article, Gosplan may well rely strongly on techniques of balancing which are not akin to supply-demand iteration, the tightening of input coefficients and the substitution of available inputs for scarce ones, for example. But I have established with the model that, in theory, at least, a basically iterative procedure of material supply planning can work.

For the record, let me note that this essay omits a number of important questions. For one thing, the model developed here does not provide for many institutional factors in Soviet-type economies which typically contribute to the inconsistencies of a plan. In addition, how capacity limitations enter into the planning procedure was not discussed. Nor was the problem of holding inventories at a reasonable level explored. These last two matters are treated in Manove (pp. 110–258), where a simulation of a version of this model and of an algorithm for nonprice rationing of intermediate goods is described.

[21] Levine (1959, p. 165). Levine reasserts this position in a later paper (1966, p. 273).

Appendix

Solution of the Model of the Soviet-Type Planning Procedure with Internal Iterations

List of Symbols:

- A = technology matrix—coefficients of input per unit output (matrix)
- Q_t = exogenously determined increments to aggregate outputs in year t (vector)
- D_t = final demands in year t (vector)
- Y_t = gross outputs in year t (vector)
- X_t = aggregate demands in year t (vector)
- E_t = supply-demand imbalances in year t (vector)
- $\hat{Y}_t$ = preliminary aggregate output targets for year t (vector)
- $\hat{X}_t$ = initial tentative aggregate demands for year t (vector)
- $Y_{t(i)}$ = tentative aggregate output targets after i internal iterations (vector)
- $X_{t(i)}$ = tentative aggregate demands after i internal iterations (vector)
- $\bar{Y}_t$ = official aggregate output targets for year t (vector)
- Ω = retrospective iteration parameter: elements pertaining to commodities retrospectively iterated are set to 1,

other elements set to 0 (diagonal matrix)

Γ = central planning parameter: elements pertaining to centrally planned commodities set to 1, other elements set to 0 (diagonal matrix)

I = identity matrix

N = number of *ex ante* iterations in the planning procedure—i.e., 1 external iteration and N-1 internal iterations (scalar)

T = the current year (scalar)

For $\Gamma^{\{n\}}$, $\Gamma^{[n]}$, $\Gamma_{\{n\}}$, $\Gamma_{[n]}$, $\Omega_{\{1\}}$, $\Omega_{[1]}$, see equations (A.9)–(A.12), (A-18), and (A.19).

Model

Tautological relationships:

$$X_t = AY_t + D_t \quad \text{(A. 1)}$$

$$E_t = Y_t - X_t \quad \text{(A. 2)}$$

Initial tentative output targets are set (a retrospective iteration):

$$\hat{Y}_t = \Omega X_{t-1} + (I - \Omega)Y_{t-1} + Q_t \quad \text{(A. 3)}$$

Aggregate demands are estimated:

$$\hat{X}_t = A\hat{Y}_t + D_t \quad \text{(A. 4)}$$

An external iteration (for centrally planned commodities only) is carried out by the center:

$$Y_{t(0)} = \Gamma\hat{X}_t + (I - \Gamma)\hat{Y}_t \quad \text{(A. 5)}$$

Internal iterations (for centrally planned commodities only) are carried out by the center. The ith internal iteration yields:

$$Y_{t(i)} = \Gamma X_{t(i-1)} + (I - \Gamma)\hat{Y}_t \quad \text{(A. 6)}$$

where

$$X_{t(i)} = AY_{t(i)} + D_t \quad \text{(A. 7)}$$

The vector of official output targets $\bar{Y}_t$ is given by $\bar{Y}_t = Y_{t(N-1)}$ where $N-1$ is the number of internal iterations that were carried out. And since it is assumed that official targets are precisely fulfilled, we have

$$Y_t = Y_{t(N-1)} \quad \text{(A. 8)}$$

Solution for the model for E_T

We define the following additional symbols:

$$\Gamma^{\{n\}} \equiv \sum_{i=0}^{n} (\Gamma A)^i \quad \text{(A. 9)}$$

$$\Gamma^{[n]} \equiv \sum_{i=0}^{n} (A\Gamma)^i \quad \text{(A. 10)}$$

$$\Gamma_{\{n\}} \equiv \Gamma^{\{n\}} - \Gamma^{\{n-1\}}\Gamma \quad \text{(A. 11)}$$

$$\Gamma_{[n]} \equiv \Gamma^{[n]} - \Gamma\Gamma^{[n-1]} \quad \text{(A. 12)}$$

It follows that:

$$A\Gamma^{\{n\}} = \Gamma^{[n]}A \quad \text{(A. 13)}$$

$$\Gamma\Gamma^{[n]} = \Gamma^{\{n\}}\Gamma \quad \text{(A. 14)}$$

$$\Gamma^{\{n\}} = I + \Gamma\Gamma^{[n-1]}A \quad \text{(A. 15)}$$

$$\Gamma^{[n]} = I + A\Gamma^{\{n-1\}}\Gamma \quad \text{(A. 16)}$$

These equivalences yield:

$$(I - A)\Gamma_{\{n\}} = \Gamma_{[n]}(I - A) \quad \text{(A. 17)}$$

We also define:

$$\Omega_{\{1\}} \equiv I - \Omega(I - A) \quad \text{(A. 18)}$$

$$\Omega_{[1]} \equiv I - (I - A)\Omega \quad \text{(A. 19)}$$

so that

$$(I - A)\Omega_{\{1\}} = \Omega_{[1]}(I - A) \quad \text{(A. 20)}$$

We can now proceed to solve equations (A.1)–(A.8) for E_T. Equations (A.4)–(A.8) yield

$$Y_t = Y_{t(N-1)} = (\Gamma A)^N \hat{Y}_t + \left(\sum_{i=1}^{N} (\Gamma A)^{N-i}\right)(\Gamma D_t + (I - \Gamma)\hat{Y}_t)$$

Applying (A.9) and (A.11) we have

$$Y_t = \Gamma_{\{N\}}\hat{Y}_t + \Gamma^{\{N-1\}}\Gamma D_t \quad \text{(A. 21)}$$

Substitution of (A.1) and (A.3) into (A.21) gives us

$$\text{(A. 22)} \quad \begin{aligned} Y_t = {} & \Gamma_{\{N\}}\Omega_{\{1\}}Y_{t-1} + \Gamma_{\{N\}}\Omega D_{t-1} \\ & + \Gamma_{\{N\}}Q_t + \Gamma^{\{N-1\}}\Gamma D_t \end{aligned}$$

We now solve for $E_t - \Gamma_{[N]}\Omega_{[1]}E_{t-1}$, which turns out to be a function only of exogenous

variables and the parameters. By (A.1) and (A.2), we have

$$E_t = (I - A)Y_t - D_t \tag{A.23}$$

Multiplying (A.23) lagged one period by $\Gamma_{[N]}\Omega_{[1]}$, and subtracting the result from (A.23) yields:

$$E_t - \Gamma_{[N]}\Omega_{[1]}E_{t-1} = (I - A)Y_t - D_t - \Gamma_{[N]}\Omega_{[1]}(I - A)Y_{t-1} + \Gamma_{[N]}\Omega_{[1]}D_{t-1}$$

and by using (A.17) and (A.20) we get

$$E_t - \Gamma_{[N]}\Omega_{[1]}E_{t-1} = (I - A)(Y_t - \Gamma_{[N]}\Omega_{[1]}Y_{t-1}) - D_t + \Gamma_{[N]}\Omega_{[1]}D_{t-1} \tag{A.24}$$

But from (A.22) we have

$$Y_t - \Gamma_{[N]}\Omega_{[1]}Y_{t-1} = \Gamma_{[N]}\Omega D_{t-1} + \Gamma_{[N]}Q_t + \Gamma^{[N-1]}\Gamma D_t$$

and substituting this into (A.24) yields, after manipulations with (A.9)–(A.20), the following:

$$E_t - \Gamma_{[N]}\Omega_{[1]}E_{t-1} = \Gamma_{[N]}(D_{t-1} - D_t) + \Gamma_{[N]}(I - A)Q_t$$

Or, defining $\Delta D_t \equiv D_t - D_{t-1}$, we have

$$E_t - \Gamma_{[N]}\Omega_{[1]}E_{t-1} = \Gamma_{[N]}(Q_t - (AQ_t + \Delta D_t)) \tag{A.25}$$

Solving (A.25) as a difference equation in t, we get

$$E_T = \sum_{t=1}^{T} (\Gamma_{[N]}\Omega_{[1]})^{T-t}\Gamma_{[N]} \cdot (Q_t - (AQ_t + \Delta D_t)) + (\Gamma_{[N]}\Omega_{[1]})^T E_0 \tag{A.26}$$

Equation (A.26) is a general solution for E_T when the planning procedure includes one external iteration and $N-1$ internal iterations. When the planning procedure includes no internal iterations ($N=1$), and given that $E_0=0$, the solution reduces to equation (14).

REFERENCES

A. Dorovskikh, "Nekotorie voprosi teori i praktiki mezhotraslevogo balansa," (Some Questions on the Theory and Practice of the Interbranch Balance), *Planovoe Khoziaistvo*, Dec. 1967, *44*, 35–44.

M. R. Eidel'man, *Mezhotraslevoi Balans Obshchestvennogo Produkta*, (The Interbranch Balance of the Social Product), Moscow 1966.

M. Ellman, "The Consistency of Soviet Plans," *Scot. J. Polit. Econ.*, Feb. 1969, *16*, 50–74.

I. A. Evenko, *Planning in the USSR*, Moscow 1962.

H. S. Levine, "The Centralized Planning of Supply in Soviet Industry," in Joint Economic Committee, U.S. Congress, *Comparisons of the United States and Soviet Economies*, Washington 1959, 151–76.

———, "Pressure and Planning in the Soviet Economy," in H. Rosovsky ed., *Industrialization in Two Systems: Essays in Honor of Alexander Gerschenkron*, New York 1966, 266–85.

———, "A Study in Economic Planning," unpublished doctoral dissertation, Harvard Univ. 1961.

M. Manove, "A Model of Administrative Planning and Plan Execution in Soviet-Type Economies," unpublished doctoral dissertation, M.I.T. 1970.

J. M. Montias, *Central Planning in Poland*, New Haven 1962.

———, "On the Consistency and Efficiency of Central Plans," *Rev. Econ. Stud.*, Oct. 1962, *29*, 280–93.

———, "Planning with Material Balances in Soviet-type Economies," *Amer. Econ. Rev.*, Dec. 1959, *49*, 963–85.

G. Sorokin, *Planning in the USSR*, Moscow 1967.

R. Stone, "Consistent Projection in Multi-Sectoral Models," in E. Mailinvaud and M. Bacharach, eds., *Activity Analysis in the Theory of Growth and Planning*, New York 1967, 232–34.

V. G. Treml, "The 1959 Soviet Input-Output Table (as Reconstructed)," in Joint Eco-

nomic Committee, U.S. Congress, *New Directions in the Soviet Economy*, Part II-A, Washington 1966, 259–70.

———, "New Soviet Interindustry Data," in Joint Economic Committee, U.S. Congress, *Soviet Economic Performance: 1966–1967*, Washington 1968, 145–58.

P. Weitzman, "Planning Consumption in the USSR," unpublished doctoral dissertation, Univ. Mich. 1969.

Transactions Costs and the Demand for Money

By Thomas R. Saving*

Much of the recent research on the demand for money treats money as essentially no different from other goods.[1] Money and goods are treated symmetrically in the budget constraint of consumers and thus, goods and money are assumed to be equally useful in transactions. This situation leads us to derive the demand for money in what is essentially a model of a transactions-costless economy. The lack of consideration of transaction cost and its effect on consumer behavior has led to rather strained explanations of why individuals use or hold money.[2] Such explanations sometimes involve arbitrary payment schedules, balanced portfolios, or perhaps simply a throwing up of the hands and saying that the utility function contains money holdings as an argument.[3] In this paper I attempt to remedy this sad state of affairs by developing the demand for money from a foundation of money's use in reducing transaction cost. The result of this approach is, I believe, both more attractive intuitively and more relevant.

I. A Pure Exchange Economy

The traditional treatment of a pure exchange economy considers individuals with utility functions of the form

$$(1) \qquad U = U(x_i); \quad i = 1, \ldots, n$$

where x_i is the rate of consumption of the ith good. Each individual receives an endowment of goods $\langle x_i^o \rangle$ which may not be carried over from period to period. Individuals are then assumed to maximize (1) subject to the constraint

$$(2) \qquad \sum_{i=1}^{n} p_i x_i^o = \sum_{i=1}^{n} p_i x_i$$

where x_i^o is the endowment of the ith good, x_i is the rate of consumption of the ith good, and p_i is the price of the ith good in terms of a base good x_b, $1 \leq b \leq n$ so that $p_b = 1$. In this model, transactions costs are zero since any good may be traded as easily as any other and trading time is either irrelevant or it is assumed that trading takes no time.

Let me now introduce transactions time as a choice variable for consumers. That is, I assume that barter transactions take time and that the consumer's time is limited and valuable to him. Thus, rewrite the utility function in (1) as

$$(3) \qquad U = U(x_i, l); \quad i = 1, \ldots, n$$

where l is the proportion of a period devoted to leisure. Additionally, let me partition each period of time into two parts: that part spent on leisure (l) and that part spent in barter (B). Thus,

$$(4) \qquad l + B = 1$$

Finally, assume that transaction time is

* Professor of economics, Texas A&M University. I am indebted to Karl Brunner, Robert Clower, Allan Meltzer, W. Phillip Gramm, and Karl Asmus for suggestions and criticisms. The research on which this article was based was financed by the National Science Foundation.

[1] Some notable exceptions are Karl Brunner and Allan Meltzer (1970) and Robert Clower.

[2] See, for example, Don Patinkin, ch. 5.

[3] See William Baumol and James Tobin for examples of fixed payments schedules. For an excellent critique of these two approaches, see Brunner and Meltzer (1967).

an increasing function of transactions undertaken:[4]

$$B = G(T), \quad \frac{\partial B}{\partial T} > 0 \tag{5}$$

where T represents the level of transactions, defined as the Euclidean length of the vector of the values of excess demands,[5]

$$T = \left\{ \sum_{i=1}^{n} [p_i(x_i - x_i^o)]^2 \right\}^{1/2} \tag{6}$$

The problem is then the maximization of (3) subject to (2) and (4). Form the following Lagrangian function

$$\begin{aligned} U^* = {} & U(x_i, l) - \lambda_1[\sum p_i(x_i - x_i^o)] \\ & - \lambda_2[l + B - 1] \end{aligned} \tag{7}$$

Differentiating (5), (6), and (7) and setting the resulting derivatives of U^* equal to zero yields the following set of necessary conditions for the maximization of (7)

$$\begin{aligned} U_i &= \left\{\lambda_1 + \lambda_2 G' \frac{p_i(x_i - x_i^o)}{T}\right\} p_i; \\ & \qquad\qquad i = 1, \ldots, n \\ U_l &= \lambda_2 \end{aligned} \tag{8}$$

[4] It can also be assumed that other resources may be substituted for time in conducting transactions. However, for my purpose nothing is gained from this complication.

[5] The Euclidean distance has several advantages. First, it is invariant with respect to changes in the units in which the x_i are measured. Second, it treats goods supplied and demanded in a symmetric fashion. Third, it is mathematically easy to work with. Note that equilibrium in the system requires that $\sum_\pi x_{i\pi} = \sum_\pi x^*_{i\pi}$ for all i, where the π are individuals, which does not imply that $T=0$. In general the equilibrium will be a repetition of the same transactions period after period. It is exactly this repetition which underlies the assertion that in a completely certain world no transactions costs could exist since individuals would simply make a once and for all infinite commitment. However, for my problem the source of the transactions cost function is irrelevant and will not be discussed further. For further elaboration of this point and an interesting discussion of the source of transactions costs, see Brunner and Meltzer (1970).

so that

$$\frac{U_i}{U_j} = \frac{\lambda_1 + U_l G' \dfrac{p_i(x_i - x_i^o)}{T}}{\lambda_1 + U_l G' \dfrac{p_j(x_j - x_j)}{T}} \frac{p_i}{p_j}; \tag{9}$$

$$i, j = , \ldots, n$$

The result indicated by equations (9) are, as expected, different from the results obtained when transactions costs are assumed to be zero. Perhaps a better idea of the difference can be had by comparing the respective equilibrium marginal rates of substitution with and without transactions costs assuming that the p_i are unaffected by their introduction. This difference is

$$\left(\frac{U_i}{U_j}\right)_T - \left(\frac{U_i}{U_j}\right)_{\sim T} \tag{10}$$

$$= \frac{U_l G'[p_i(x_i - x_i^o) - p_j(x_j - x_j^o)]}{T\lambda_1 + U_l G' p_j(x_j - x_j^o)} \frac{p_i}{p_j}$$

where $(U_i/U_j)_T$ and $(U_i/U_j)_{\sim T}$ represents the case with and without transactions costs, respectively. Since λ, T, U_l, G', and the p_i are all positive, the sign of (10) depends on the excess demands. In particular, if the consumer is a net supplier of good i, i.e., $(x_i - x_i^o) < 0$, and a net demander of good j, then $(U_i/U_j)_{\sim T} > (U_i/U_j)_T$. In a two-good world this would require an increase in the consumption of good i relative to good j.

This result is intuitively obvious once the effect of the introduction of transactions costs on the prices of the goods is recognized. Essentially, what has happened is that the price received for goods supplied has fallen in the sense that while the same amount of any other good is obtained the transaction now involves a positive expenditure of time. In contrast, the price paid for goods purchased has

risen since in addition to the former amount of other goods given up some transactions time must be expended. Thus, it is not surprising that the introduction of transactions costs results in increased consumption of supplied goods (taken as a group) and decreased consumption of demanded goods (again, taken as a group).

The effect of introducing transaction cost is illustrated graphically in Figure 1. In the figure the vertical axis represents time per period, which is necessarily unity. The right-hand axis measures units of good 1 and the left-hand axis measures units of good 2. Point A in Figure 1 is the initial endowment of goods and time $(x_1^o, x_2^o, 1)$. Line CC is the budget line for the given prices (p_1, p_2); it represents the relevant boundary of attainable combinations of x_1, x_2, and l under conditions of no transaction cost and non-satiety for x_1, x_2, and l.

The effect of transaction cost is to reduce the height of CC whenever the desired combination of (x_1, x_2) moves away from (x_1^o, x_2^o). This effect is shown in the figure by the movement of the boundary from CC to $C'C'$. Thus allowing for transaction cost reduces the set of attainable combinations of (x_1, x_2, l) by the sum of the two areas (CAC') to the right and left of A. If money is adopted by a community in which individuals are characterized by the situation in Figure 1, this money must have the effect of raising the boundary $C'C'$ (e.g., from $C'C'$ to $C''C''$ in Figure 1).[6]

II. A Money Economy

The essential difference between money and non-money economies is that in a money economy at least one good is universally accepted in transactions so that

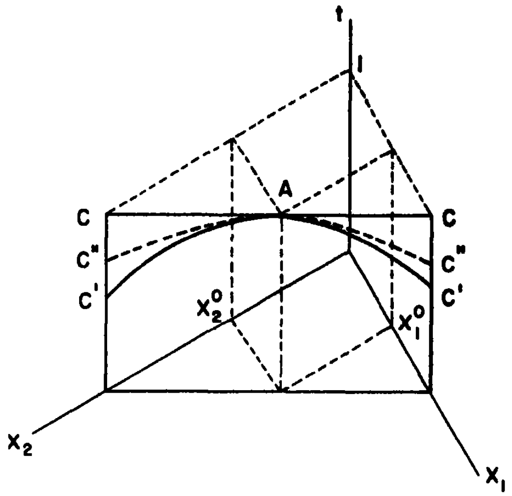


FIGURE 1

the number of transactions involved in any trade is at most two; one trade of a supplied item for money and then a trade of money for a desired item. However, the existence of money does not preclude direct barter whenever such barter is less time consuming than two money exchanges. For example, if one desires to exchange two hours work for a transistor radio, he may sell the labor for money and take the money and trade it for a transistor radio, or if he knows someone who will take two hours of labor for a transistor radio then the trade can be made directly. Moreover, in the latter case having to use money may involve both traders in an extra burden. Thus, universal acceptance of money is interpreted to mean that having money is a sufficient condition for exchange but not necessary.

One obvious use of money, then, is the reduction of exchange time for the members of the community. In this use, money is used until the marginal cost of using it is equal to the marginal value of time saved through its use. However, money frequently performs other functions such as a convenient store of value; meaning

[6] If it were assumed that in addition to time, resources may be used in transactions, then the effect on Figure 1 would be to convert the relevant boundary to a surface rising from line $C'C'$ in the direction of less x_1 or less x_2.

that its storage costs are lower than the storage costs of other goods and services.[7] Thus, the institution of money allows one to avoid storage costs on some commodities, first by making transactions costs lower so that transactions may be engaged in more frequently and secondly, by allowing the separation of the sale of supplied goods from the purchase of goods for consumption.

Consider an individual who desires to maximize a utility function of the following form

$$(11)\quad U = U(c_t, l_t);\quad t = 0, \ldots, H$$

where l_t and c_t, respectively, are the rates of consumption of leisure, and goods and services in the period t, and H is the end of the consumers horizon. Assume additionally that the individual faces income and time constraints

$$(12)\quad y_t = H(w_t, t);\ H_{w_t} > 0 \text{ and } H_{w_t w_t} = 0$$

and

$$(13)\qquad l_t + w_t + T_t = 1$$

where y_t is real income in period t, w_t is work time in period t and T_t is the rate of transactions time expenditure in period t.[8] Equation (13) indicates that for each interval the sum of the three rates of spending time must exhaust the interval and it is additionally assumed that the ways to spend time are mutually exclusive.

Any given levels of y_t and c_t imply an aggregate level of transactions. These transactions may be carried out all at once, sporadically, or regularly throughout the interval. Additionally, the transactions may be consummated via direct barter or with a media of exchange. For every distribution of transactions and time path of income and consumption over the interval, there will exist unique levels of average stocks of work services delivered but not collected for, $\bar{y}_t$, consumption goods, $\bar{c}_t$, and real holdings of the various media of exchange, $\bar{m}_t^i$ $(i=1, \ldots, n)$, e.g., currency and demand deposits. Subject to the costs of holding these inventories, consumers will select the distribution of transactions so as to minimize transactions time. I shall assume the following transactions time function[9]

$$(14)\quad T_t = G(y_t, c_t, \bar{y}_t, \bar{c}_t, \bar{m}_t^i);\quad i = 1, \ldots, n$$

where

$$(15)\quad \begin{matrix} G_{y_t}>0,\ G_{c_t}>0,\ G_{\bar{y}_t}<0,\ G_{\bar{c}_t}<0,\ G_{\bar{m}_t^i}<0; \\ t = 0, \ldots, H;\quad i = 1, \ldots, n \end{matrix}$$

Since the remainder of this discussion follows directly from assumptions concerning (14) further discussion of it is warranted. First I shall assume that time required for transactions is dependent on both the number and size of the transaction engaged in. Thus, a doubling of income and consumption if handled by doubling the size of each transaction will result in a doubling of $\bar{y}_t$, $\bar{c}_t$, and $\bar{m}_t^i$ and with a resultant increase in transactions costs.[10] The signs of the partial derivatives of (14) (shown as (15)) result from the following: 1) an increase in income must increase transactions time if no change occurs in the other arguments because in this case the only way to sell the increased output is to increase the number of transactions;[11] 2) an increase in consumption increases transactions costs for the same

[7] In the context of this paper these storage costs are direct costs since I will deal with a world in which future prices are known to be equal to present prices with certainty.

[8] H_{w_t} and $H_{w_t w_t}$ are the first-and second-order partial derivatives of H. Throughout the remainder of the paper the various functions will be subscripted in a a similar fashion to denote partial differentiation.

[9] Thus, (14) shows the minimum transactions time attainable for each level of the arguments.

[10] Note that if transactions time is independent of the size of transactions but depends only on the number of transactions engaged in, then (14) will be homogeneous of degree zero.

[11] Simply increasing the size of transactions will increase $\bar{y}_t$ and $\bar{m}_t$, which is, of course, not allowed.

reason as increases in income; 3) an increase in the average stock of produced goods and services implies a reduction in the number of sale transactions (given y_t) and thus a reduction in transactions time; (4) an increase in the average stock of consumed goods and services implies a reduction in the number of purchase transactions (given c_t) and thus a reduction in transactions time; 5) an increase in the average stock of the ith type of money implies (a) larger sale and purchase transactions (given y_t and c_t) and thus fewer of them or (b) that a greater proportion of all transactions are money transactions. Both (a) and (b) above imply a reduction in transaction time.[12]

Assume additionally that each stock has diminishing returns and that income and consumption have increasing transactions cost so that

$$G_{yy} > 0,\ G_{cc} > 0,\ G_{\bar{y}\bar{y}} > 0,\quad G_{\bar{c}\bar{c}} > 0,\ G_{ii} > 0 \quad \text{for all } i \tag{16}$$

and that the only other assets besides the stocks of produced goods and services, consumed goods and services, and the various types of money, are bonds yielding interest rate r per period. Thus, the cost of holding the stocks, $\bar{y}$, $\bar{c}$, $\bar{m}^i$ is the income foregone on bonds that could have been held plus any additional holding costs.[13] Finally, assume that each individual faces a budget constraint of the form shown in equation (17). Note that r is the rate of interest, b is initial real bond holdings, m is initial real money holdings, $\delta_{\bar{y}}$, $\delta_{\bar{c}}$, and the $\delta_{\bar{m}}{}^i$ are the direct costs of holding in-

$$\sum_{t=0}^{H}\left(\frac{1}{1+r}\right)^t y_t + b + m = \sum_{t=0}^{H}\left(\frac{1}{1+r}\right)^t \left[c_t + r\left(\bar{y}_t + \bar{c}_t + \sum_i \bar{m}_t^i\right) + \delta_{\bar{y}_t}\bar{y}_t + \delta_{\bar{c}_t}\bar{c}_t + \sum_i \delta_{\bar{m}_t^i}\bar{m}_t^i\right] \tag{17}$$

ventories of produced goods, consumer goods, and real money, respectively. Additionally, assume that $\delta_{\bar{c}} \geq 0$, $\delta_{\bar{y}} \geq 0$ and $\delta_{\bar{m}}{}^i < \delta_{\bar{c}}$, $\delta_{\bar{y}}$ for all i, and finally that if for any i, $\delta_{\bar{m}}{}^i < 0$ then $r \geq |\delta_{\bar{m}}{}^i|$.[14] The term $r(\bar{y}_t + \bar{c}_t + \sum_i \bar{m}_t^i)$ on the right-hand side of (17) is necessary because the left-hand side of (17) assumes that all of y_t, b, and m is invested at rate of interest r in arriving at the present value. Thus, to the extent that the consumer holds stocks of output, consumption goods, or money, he gives up the return assumed by the left-hand side of (17).

Since the problem of interest here is the demand for money, the determinants of the optimal paths of consumption and leisure ($\hat{c}_t$ and $\hat{l}_t$, respectively) may be assumed to have already been chosen; that is, they may be treated as parameters.[15] It can readily be shown that a

[12] Note that in a certainty world, the minimum $\bar{c}$, $\bar{y}$, $\bar{m}^i$ in each period will be zero.

[13] If the assumption of constant goods prices is dropped and replaced by a known rate of inflation then this rate of inflation may be treated as an additional holding cost of money. In this approach the bonds become fixed in real value so that r represents a real rate of interest. The money rate of interest would then be $r+\rho$, where ρ is the known rate of change in prices. The holding costs of the various forms of money would be $\delta_{\bar{m}}{}^i + \rho$ provided that all forms of money are suffering the same rate of price level depreciation.

[14] Equilibrium for the economic system with a positive level of capital will require that $r > |\delta_{\bar{m}}{}^i|$ if $\delta_{\bar{m}}{}^i < 0$, for all i. Additionally, equilibrium in the banking system will also require the same condition.

[15] This approach can be viewed as a sort of inverse "indirect utility" approach in that instead of the substitution of demand functions into the utility function, I have substituted certain arguments of the utility function into the constraint, i.e., an indirect wealth maximization approach. In this approach the consumer is viewed as a producer and consumer of wealth and given the amount he desires to consume, he maximizes the output of wealth. Thus, I have treated this problem in a way similar to the minimization of cost for given output in the theory of the firm. where even though only one output is the profit maximizing one, it is true that

necessary condition for utility maximization is that for given $\langle \hat{c}_t, \hat{l}_t \rangle$ the levels of $\bar{y}_t$, $\bar{c}_t$, and the $\bar{m}_t^i$ be chosen so that the following implicit form of (17) be a maximum

$$W^* = \sum_t \left(\frac{1}{1+r}\right)^t y_t + b + m - \sum_t \left(\frac{1}{1+r}\right)^t \cdot \left[c_t + r\left(\bar{y}_t + \bar{c}_t + \sum_i \bar{m}_t^i\right) + \delta_{\bar{y}}\bar{y}_t + \delta_{\bar{c}}\bar{c}_t + \sum_i \delta_{\bar{m}^i}\bar{m}_t^i\right] \tag{18}$$

subject to (12), (13), and (14).[16]

Now substituting (12) and (13) into (18) and (14) yields the following function to be maximized

$$W^* = \sum_t \left(\frac{1}{1+r}\right)^t H[1-(\hat{l}_t + T_t), t] + b + m - \sum_t \left(\frac{1}{1+r}\right)^t \cdot \left[\hat{c}_t + r\left(\bar{y}_t + \bar{c}_t + \sum_i \bar{m}_t^i\right) + \delta_{\bar{y}}\bar{y}_t + \delta_{\bar{c}}\bar{c}_t + \sum_i \delta_{\bar{m}^i} m_t^i\right] \tag{19}$$

subject to

$$T = G\{H[1-(\hat{l}_t + T)], \hat{c}_t, \bar{y}_t, \bar{c}_t, \bar{m}_t^i\} \tag{20}$$

Differentiating (19) subject to (20) and setting the result equal to zero yields the first-order conditions for a maximum of (19)[17]

$$-\left(\frac{H_{wt}G_{\bar{y}_t}}{1+G_{y_t}H_{wt}}\right) = (r + \delta_{\bar{y}})$$
$$-\left(\frac{H_{wt}G_{\bar{c}_t}}{1+G_{y_t}H_{wt}}\right) = (r + \delta_{\bar{c}}) \qquad t=0, \ldots, H \tag{21}$$
$$-\left(\frac{H_{wt}G_{\bar{m}_t^i}}{1+G_{y_t}H_{wt}}\right) = (r + \delta_{\bar{m}^i}) \qquad i = 1, \ldots, n$$

where the terms on the left-hand side of equations (21) are the marginal return to holding average stocks of goods and services produced, goods and services consumed, and money balances, respectively.

for any given output costs must be a minimum. In my case this boils down to for any given level of c and l wealth must be a maximum. In particular at the utility maximizing levels of c and l wealth must be a maximum and this maximum is, of course, zero. Essentially my approach says that since (18) must be a maximum for any levels of c and l chosen, (18) must be a maximum for the particular levels of c and l that maximize utility. Treating the problem in this way has the distinct advantage of simplifying the mathematics and of making the demand for money a function of the levels of choices concerning consumption and leisure.

[16] If W^* is not a maximum then some other values of $\bar{y}_t$, $\bar{c}_t$, $\bar{m}$ must be possible which will result in the left-hand side of (18) being greater than the right-hand side for $(\hat{c}_t, \hat{l}_t)$ allowing positive increments to c_t and l_t and therefore an increase in utility. Note that if $(\hat{c}_t, \hat{l}_t)$ is optimal then max $W^*=0$, but is nonetheless a maximum.

[17] Had I treated the problem as one of utility maximization equations (21) would still have been necessary conditions for a maximum. In addition, the following two conditions would also have been necessary

$$U_{c_t} = -\lambda\left(\frac{1}{1+r}\right)^t\left[\frac{1+H_w(G_y+G_c)}{1+G_yH_w}\right] \qquad t=0, \ldots, H \tag{21*}$$
$$U_{l_t} = -\lambda\left(\frac{1}{1+r}\right)^t\left[\frac{H_w}{1+G_yH_w}\right]$$

Moreover, the conditions required for wealth maximization would still be necessary to derive the derivatives of the demand functions presented as equations (22). Essentially, the approach used here derives a wealth function.

$$W = g(c_t, l_t; r, \delta_{\bar{y}}, \delta_{\bar{c}}, \delta_{\bar{m}^i}) \tag{a}$$

which shows the maximum wealth for each set of values of the arguments. Utility maximization then requires that the following Lagrangian be a maximum

$$U^* = U(c_t, l_t) - \lambda[g(c_t, l_t; r, \delta_{\bar{y}}, \delta_{\bar{c}}, \delta_{\bar{m}^i})] \tag{b}$$

where the function (a) is required to be equal to zero. Thus, the similarity between the usual treatment of profit maximization and the problem considered here is now quite distinct

The denominators in (21) reflect the fact that the transactions time saved cannot be entirely converted into income for fixed stocks, because increased income requires increased transactions time.

Equations (21) plus the initial conditions (12), (13), and (14) make up, for each t, $(n+5)$ equations in the $(n+5)$ unknowns T_t, w_t, y_t, $\bar{y}_t$, $\bar{c}_t$, $\bar{m}_t^i$ and parameters $\hat{c}_t$, $\hat{l}_t$, r, $\delta_{\bar{y}}$, $\delta_{\bar{c}}$, $\delta_{\bar{m}^i}$. The system, (21), can be solved for each of the variables in terms of the parameters.[18] Let me write the solutions for $\bar{y}_t$, $\bar{c}_t$, $\bar{m}_t^i$ as[19]

$$\hat{\bar{m}}_t^i = \mu^i(\hat{c}_t, \hat{l}_t, r, \delta_{\bar{y}}, \delta_{\bar{c}}, \delta_{\bar{m}}^j, t)$$

$$(22)\quad \hat{\bar{c}} = \gamma(\hat{c}_t, \hat{l}_t, r, \delta_{\bar{y}}, \delta_{\bar{c}}, \delta_{\bar{m}}^j, t); \qquad i, j = 1, \ldots, n$$

$$\hat{\bar{y}}_t = \phi(\hat{c}_t, \hat{l}_t, r, \delta_{\bar{y}}, \delta_{\bar{c}}, \delta_{\bar{m}}^j, t)$$

The derivatives of functions (22) can be derived by totally differentiating system (21), (12), (13), and (14). After substitution of (12), (13), and (14) and assuming that the cross-partials involving y and c are zero, the differentiation results in the system of differential equations shown as (23). For expositional convenience I have dropped the time subscripts, and ordered the stocks of the n types of money, produced goods, and consumed goods so that the types of money are denoted as σ_1, σ_2,

$$(23)\quad \sum_{j=1}^{n+2}[G_{ij} + (r+\delta_i)(r+\delta_j)G_{yy}]d\sigma_j = (r+\delta_i)\left(\frac{H_wG_{yy}}{1+H_wG_y}\right)(d\hat{l} + G_c d\hat{c}) - \left(\frac{1+G_yH_w}{H_w}\right)(dr + d\delta_i); \qquad i = 1, \ldots, n+2$$

$\ldots$, σ_n, produced goods as σ_{n+1}, consumed goods σ_{n+2}, $\delta_{\bar{m}^i}$ as δ_i $(i=1, \ldots, n)$, $\delta_{\bar{y}}$ as δ_{n+1}, and $\delta_{\bar{c}}$ as δ_{n+2}.

Denote the matrix of coefficients of system (23) as

$$(24)\quad S = [G_{ij} + (r+\delta_i)(r+\delta_j)G_{yy}]; \qquad i, j = 1, \ldots, n+2$$

then the solution for $d\sigma_j$ is

$$(25)\quad d\sigma_j = \sum_{i=1}^{n+2}(r+\delta_1)\frac{H_wG_{yy}}{1+H_wG_y}\frac{S_{ij}}{|S|}\cdot(d\hat{l}+G_c d\hat{c}) - \sum_{i=1}^{n+2}\frac{1+G_yH_w}{H_w}\frac{S_{ij}}{|S|}\cdot(dr + d\delta_i);$$

where $|S|$ is the determinant of S and S_{ij} is the cofactor of S associated with the element common to the ith row and jth column. The second-order conditions for a maximum of (18) require that

$$(26)\quad T = -\frac{H_w}{(1+H_wG_y)}S$$

be negative definite. Thus, S itself must be positive definite. This information allows the sign of the own partials of the demand functions (22) to be determined, i.e.,

$$(27)\quad \frac{\partial\sigma_j}{\partial\delta_j} = -\frac{1+G_yH_w}{H_w}\frac{S_{jj}}{|S|} < 0$$

since $|S|$ and $S_{jj} > 0$ for all j, but leaves the signs of all other partials undetermined.

However, if it is assumed that the effect of a change in the cost of holding the jth

[18] This follows from the fact that equations (12), (13), (14) are monotonic in T_t, w_t, and y_t, and can be solved uniquely for T_t, w_t, and y_t in terms of the $n+2$ remaining variables. Then these solutions may be substituted into (21). The positive definiteness of S, the Jacobian of the system, then insures the existence of a unique solution of the $n+2$ variables in terms of the parameters.

[19] Note that income and assets do not appear directly in the demand for money function. The reasons for this are twofold. First, income is an endogeneous variable and thus on the same footing as money. However, the income function is assumed fixed (as is the transactions cost function) and will of course influence the demand for money. Second, initial assets affect the demand for money indirectly through c_t, but for our problem c_t is fixed.

stock on the demand for the jth stock is greater than the sum of the effects of changes in all other holdings costs on the demand for the jth stock, i.e.,

$$(28)\qquad \left|\frac{\partial\sigma_j}{\partial\delta_j}\right| > \sum_{i\neq j}^{n+2}\left|\frac{\partial\sigma_j}{\partial\delta_i}\right|,$$

then much more can be said.[20] In particular,

$$(29)\qquad \frac{\partial\sigma_j}{\partial r} = -\sum_{i=1}^{n+2}\frac{1+G_yH_w}{H_w}\,\frac{S_{ij}}{|S|} < 0 \qquad \text{for all } j$$

since the sum on the right-hand side of (29) is now dominated by the term $S_{jj}/|S|$ which is positive. In addition, while the effect of a change in desired leisure or consumption on any particular stock is still indeterminate the effect on the sum of all stocks is now determinate and is

$$(30)\qquad \begin{aligned}\frac{\partial(\sum\sigma_i)}{\partial\hat{l}} &= \sum_{i=1}^{n+2}(r+\delta_i)\frac{H_wG_{yy}}{1+H_wG_y}\cdot\sum_{j=1}^{n+2}\frac{S_{ij}}{|S|} > 0\\ \frac{\partial(\sum\sigma_i)}{\partial\hat{c}} &= \sum_{i=1}^{n+2}(r+\delta_i)G_c\frac{H_wG_{yy}}{1+H_wG_y}\cdot\sum_{j=1}^{n+2}\frac{S_{ij}}{|S|} > 0\end{aligned}$$

This result follows because (28) implies that $\sum_{j=1}^{n+2} S_{ij}/|S| > 0$ for all i so that every term on the right-hand side of equation (30) is positive.

Finally some stronger results can be had if it is assumed that $G_{ij}=0$ for all i and j, $i\neq j$. For in this case it can be established through a series of elementary row and column operations on S that the cofactors of S are[21]

$$(31)\qquad \begin{aligned}S_{ij} &= -(r+\delta_i)(r+\delta_j)G_{yy}\prod_{k\neq i,j}^{n+2}G_{kk} < 0; \qquad i\neq j\\ S_{jj} &= \prod_{k\neq j}^{n+2}G_{kk} + G_{yy}\sum_{i\neq j}^{n+2}(r+\delta_i)^2\cdot\prod_{k\neq i,j}^{n+2}G_{kk} > 0\end{aligned}$$

From (25) and (31) it follows that: $\partial\sigma_j/\partial\delta_i>0$ for all $j\neq i$; $\partial\sigma_j/\partial\delta_j<0$, $\partial\sigma_j/\partial\hat{l}>0$, $\partial\sigma_j/\partial\hat{c}>0$ for all j; $\partial\sigma_j/\partial r \gtrless 0$. Thus, for this case, increases in desired leisure or consumption increase the average level of all stocks, both money and goods. Increases in the storage costs of a stock will decrease the average level of that stock and increase the average levels of all other stocks.[22] Finally, increases in the interest rate may increase, leave unchanged, or decrease a particular average stock depending on that stock's storage costs relative to the storage costs of other stocks. Hence, increases in the rate of interest will not necessarily reduce average holdings of each type of money. However, such increases will reduce the average total value of all stocks held as can be seen from equation (32).

In addition, if the storage costs on all forms of money are less than the storage costs on both produced and consumed goods then an increase in the interest rate will necessarily reduce the average stock of the sum of all monies. This is shown in

[20] This is equivalent to assuming that S^{-1} has a dominant diagonal since the $S_{ij}/|S|$ are the elements of S^{-1}.

[21] It can be shown that for any (nxn) matrix $A=(a_{ij})$ where $|a_{ii}|>|a_{ij}|$ for all i and j, $i\neq j$, $a_{ii}>0$ for all i, $a_{ij}=a_{lk}$ for all i, j, l, k, $i\neq j$, $l\neq k$ than a matrix $B=(b_{ij})$ $=A^{-1}$ exists and $b_{ii}>0$ for all i, $b_{ij}<0$ for all i and j, $i\neq j$. Thus, many of the following results will hold if $G_{ij}/(r+\delta_i)(r+\delta_j)=G_{lk}/(r+\delta_l)(r+\delta_k)$ for all i, j, k, l such that $i\neq j$, $k\neq l$. Of course, $G_{ij}=0$ for all i, j, $i\neq j$ is a special case of this condition.

[22] Since a known rate of inflation may be treated as a storage cost of holding money, an increase in the rate of inflation is equivalent to an increase in storage costs of money. Thus, such an increase will reduce the average level of money balances and increase the level of stocks of goods.

$$(32)\quad \frac{\partial(\sum \sigma_j)}{\partial r} = -\left(\frac{1 + G_y H_w}{H_w}\right) \cdot \left\{ \sum_{j=1}^{n+2} \prod_{k \neq j}^{n+2} G_{kk} + G_{yy} \sum_{i,j=1}^{n+2} \cdot [(r + \delta_i) - (r + \delta_j)]^2 \cdot \prod_{k \neq i,j}^{n+2} G_{kk} \right\} < 0$$

equation (33) which is necessarily less than zero if $\delta_j < \delta_{n+1}, \delta_{n+2}$ for all $j \leq n$. Thus, while the value of the entire money portfolio must be negatively related to the interest rate, no individual element in the portfolio need be so related. The reason for this seemingly strange result is that changes in the rate of interest do not leave the relative marginal costs of holding the various stocks unchanged. In particular, consider the marginal cost of holding stocks of ith money type relative to the marginal cost of holding stocks of the jth money

$$(34)\quad \frac{MC_i}{MC_j} = \frac{r + \delta_i}{r + \delta_j}$$

The derivative of (34) with respect to the interest rate is

$$(35)\quad \frac{\partial\left(\frac{MC_i}{MC_j}\right)}{\partial r} = \frac{\delta_j - \delta_i}{(r + \delta_j)^2}$$

Thus, if $\delta_j < \delta_i$, increases in the interest rate reduce the marginal cost of the ith money relative to the jth money, and substitution toward the ith money form will occur.[23] This substitution effect may be large enough so that an increase in holdings of the ith money form result. Note that if the holding costs of money are always less than the holding costs of goods then an increase in interest rates will result in decreases in the price of holding goods relative to any money form.

III. An Alternative Approach

Some additional insight into the nature of the results can be had by viewing the problem as one of the minimization of transactions time for each period τ subject to the condition that (18) equal zero where $\hat{l}_t$, $\hat{w}_t$, $\hat{c}_t$, $(t=0, \ldots, H)$ and the σ_{jt} $(t=0 \ldots, \tau-1, \tau+1, \ldots, H)$ are treated as parameters. Such a minimization of transactions time is a necessary condition

[23] In a model with a known rate of inflation ρ, the marginal cost of the ith money type relative to the jth money type is

$$(34^*)\quad \frac{MC_i}{MC_j} = \frac{r + \rho + \delta_i}{r + \rho + \delta_j}$$

and the derivative with respect to the rate of inflation is

$$\frac{\partial\left(\frac{MC_i}{MC_j}\right)}{\partial \rho} = \frac{\delta_j - \delta_i}{(r + \rho + \delta_j)^2}$$

Thus, if $\delta_j < \delta_i$ an increase in the known rate of inflation reduces the marginal cost of the ith money relative to the jth money.

$$(33)\quad \frac{\partial\left(\sum^{n} \sigma_j\right)}{\partial r} = -\left(\frac{1 + G_y H_w}{H_w}\right) \sum_{j=1}^{n} \left\{ \prod_{k \neq j}^{n+2} G_{kk} + G_{yy} \sum_{i,j=1}^{n} [(r + \delta_i) - (r + \delta_j)]^2 \cdot \prod_{k \neq i,j}^{n+2} G_{kk} + G_{yy}(r + \delta_{n+1}) [(r + \delta_{n+1}) - (r + \delta_j)] \cdot \prod_{k \neq n+1,j}^{n+2} G_{kk} + G_{yy}(r + \delta_{n+2}) \cdot [(r + \delta_{n+2}) - (r + \delta_j)] \cdot \prod_{k \neq j}^{n+1} G_{kk} \right\} < 0$$

for utility maximization since, if transactions time in period τ could be reduced while keeping $\hat{c}_t$, $\hat{l}_t$, $\hat{w}_t$ (t=0, . . . , H) and σ_{jt} ($t \neq \tau$) constant, then this additional time could be used to increase $\hat{l}_t$ or $\hat{c}_t$ or both for some or all t and thus increase utility. Viewed in this way the problem is to minimize

$$(36) \quad T_\tau = G(\mathcal{J}_\tau, \hat{c}_\tau, \sigma_{j\tau}); \quad j = 1, \ldots, n+2$$

subject to

$$(37) \quad (1+r)^\tau \left\{ \sum_{t=0}^{H} \left(\frac{1}{1+r}\right)^t (\mathcal{J}_t - \hat{c}_t) + b + m - \sum_{t \neq \tau}^{H} \left(\frac{1}{1+r}\right)^t \sum_{i=1}^{n+2} (r+\delta_i)\sigma_{it} \right\} = \sum_{i=1}^{n+2} (r+\delta_i)\sigma_{i\tau}$$

Now form the Lagrangian function

$$(38) \quad T_\tau^* = G - \lambda \left[A - \sum_{i=1}^{n+2} (r+\delta_i)\sigma_{i\tau} \right]$$

where A is the left-hand side of (37).[24] The first-order conditions for the minimization of (38) are

$$(39) \quad G_i = -\lambda(r+\delta_i); \quad i = 1, \ldots, n+2$$

which may be expressed as

$$(40) \quad \frac{G_i}{G_j} = \frac{(r+\delta_i)}{(r+\delta_j)} \quad \text{for all } i \text{ and } j.^{25}$$

From (40) then, the ratios of the marginal costs of holding the various stocks must equal the ratios of their respective marginal transactions time returns.[26] The second-order conditions require that the quadratic form $Q = \sum\sum_{ij} G_{ij} d\sigma_i d\sigma_j$ be positive definite subject to the condition that $\sum_{i=1}^{n+2} (r+\delta_i)\, d\sigma_i = 0$. Thus, the following bordered Hessian must be positive definite.

$$(41) \quad T = \begin{bmatrix} 0 & (r+\delta_j) \\ (r+\delta_i) & G_{ij} \end{bmatrix}; \qquad |T| < 0,\ T_{ii} < 0 \quad \text{for all } i$$

where $|T|$ is the determinant of matrix T and the T_{ii} are the principal minors of matrix T. For each t, equations (39) plus equation (37) comprise ($n+3$) equations in the ($n+3$) unknowns (σ_i, λ) and can be solved for the unknowns in terms of the parameters (A, r, δ_i) so that the solution for σ_j may be expressed as[27]

$$(42) \quad \sigma_j = S_i(A, r, \delta_i); \quad i = 1, \ldots, n+2$$

Now the signs of the partial derivatives of equations (42) may be determined by totally differentiating the equation system consisting of (39) and (37) which yields

$$(43) \quad \sum_{j=1}^{n+2} (r+\delta_j) d\sigma_{j\tau} = dA + \sum_{j=1}^{n+2} \sigma_{j\tau}(dr + d\delta_j)$$

$$(r+\delta_i)d\lambda + \sum_{j=1}^{n+2} G_{ij} d\sigma_{j\tau} = -\lambda(dr + d\delta_i); \quad i = 1, \ldots, n+2$$

The matrix of coefficients of (43) is exactly T. Therefore, the general solution of (43) may be written as

$$(44) \quad d\sigma_{j\tau} = \frac{T_{oj}}{|T|}\left[dA + \sum_{i=1}^{n+2} \sigma_{i\tau}(dr + d\delta_i) \right] - \lambda \sum_{i=1}^{n+2} \frac{T_{ij}}{|T|}(dr + d\delta_i)$$

[24] Note that the level of resources available for transactions is treated as a parameter in its entirety for this problem. In a general utility maximizing model this is, of course, inappropriate.

[25] Since the choice of τ is arbitrary (39) must hold for all τ so that (40) can be written as

$$(40^*) \quad \frac{G_{it}}{G_{j,t+k}} = \frac{\left(\frac{1}{1+r}\right)^t (r+\delta_i)}{\left(\frac{1}{1+r}\right)^{t+k} (r+\delta_j)} \quad \text{for all } i, j, \text{ and } k$$

[26] Note the ratios of the marginal transactions time returns equals the ratios of the marginal revenues since the transactions time saved is valued at H_w no matter how it is saved.

[27] The nonsingularity of T, the Jacobian of the system, guarantees the uniqueness of equation (42).

so that the partial derivatives in which I am interested are[28]

$$\frac{\partial \sigma_{jr}}{\partial A} = \frac{T_{oj}}{|T|}$$

$$(45)\qquad \frac{\partial \sigma_{jr}}{\partial \delta_i} = \sigma_{ir}\frac{T_{oj}}{|T|} - \lambda \frac{T_{ij}}{|T|}\,; \qquad i, j = 1, \ldots, n+2$$

$$\frac{\partial \sigma_{jr}}{\partial r} = -\sum_{i=1}^{n+2} \sigma_{ir}\frac{T_{oj}}{|T|} - \lambda \sum_{i=1}^{n+2} \frac{T_{ij}}{|T|}$$

Using this approach $\partial\sigma_j/\partial A$ may be viewed as the resource effect, or since the change in A must be the result of a decision to change transactions time in period r, as a transactions time effect. In this way, $\partial\sigma_j/\partial\delta_i$ may be written as

$$(46)\qquad \frac{\partial \sigma_{jr}}{\partial \delta_i} = \sigma_{ir}\frac{\partial \sigma_j}{\partial A} + M_{ij}$$

where $M_{ij} = -\lambda T_{ij}/|T|$ is interpreted as the substitution effect. That is, M_{ij} is the effect of a change in the costs of holding the ith stock on the average holdings of the jth stock with transactions time constant. Note here that if we assume as earlier that $G_{ii} > 0$ for all i and $G_{ij} = 0$ for $i \neq j$ then

$$T_{oj} = -(r + \delta_j)\prod_{k\neq j}^{n+2} G_{kk} < 0$$

$$(47)\qquad T_{ij} = (r + \delta_i)(r + \delta_j)\prod_{k\neq i,j}^{n+2} G_{kk} > 0$$

$$T_{jj} = \sum_{i\neq j}^{n+2} \frac{(r + \delta_i)^2}{G_{ii}} \prod_{k=1}^{n+2} G_{kk} < 0$$

In this case the signs of the partial derivatives are $\partial\sigma_j/\partial A > 0$ for all j; $\partial\sigma_j/\partial\delta_i > 0$ for all i and j, $i \neq j$; $\partial\sigma_j/\partial\delta_j < 0$ for all j; and $\partial\sigma_j/\partial r \gtrless 0$; depending on the relation between the costs of holding σ_j relative to the σ_i $(i \neq j)$.

Graphically the analysis is depicted in Figure 2 where the vertical axis represents the average stock of σ_i and the horizontal axis represents the average stock of σ_j. The iso-transactions time loci are labeled T^{-1}, T^0, and T^1 and are drawn convex to the origin as required by (41). The constraint is then represented by the linear function with intercepts at

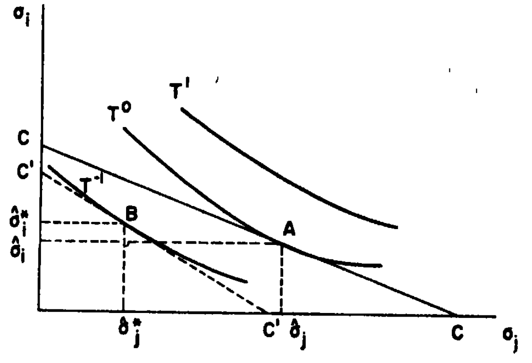


FIGURE 2

$$\sigma_i = \frac{A - \sum_{k\neq i,j} (r + \delta_k)\sigma_k}{(r + \delta_i)}$$

and

$$\sigma_j = \frac{A - \sum_{k\neq i,j} (r + \delta_k)\sigma_k}{(r + \delta_j)},$$

respectively. The figure is constructed so that $\delta_j < \delta_i$ and the slope of the constraint is accordingly less than one in absolute value. Point A in the figure is the transactions time minimizing combination of σ_i, σ_j given the $\hat{\sigma}_k$ $(k \neq i, j)$. Thus, in this way the vector of optimal stocks is determined. This vector of optimal stocks includes, of course, all the optimal stocks of the various forms of money $\bar{m}^i$ $(i = 1, \ldots, n)$ and thus the portfolio of money forms is determined.

Assume that σ_i and σ_j are both forms of money and allow an increase in the interest rate. As I have shown in (35) such a change will not leave the slope of the

[28] Note that the partials with respect to the δ_i and r are compensated partials. That is, it is assumed that the consumer is compensated so that A remains unchanged.

constraint unchanged. In fact,

$$\frac{\partial\left(\frac{d\sigma_i}{d\sigma_j}\right)}{\partial r} = \frac{\partial\left[\frac{(r+\delta_j)}{(r+\delta_i)}\right]}{\partial r} \tag{48}$$

$$= -\left[\frac{(r+\delta_i)-(r+\delta_j)}{(r+\delta_i)^2}\right]$$

$$= -\frac{(\delta_i-\delta_j)}{(r+\delta_i)^2}$$

so that relative prices change in favor of the more expensive stock when the rate of interest rises. Therefore, it is possible that the substitution effect toward the more expensive money will dominate the resource effect and the increase in the interest rate will increase the stocks of the more expensive money. (Point B in Figure 1 is an example of such a case.) Given a transactions time function that is homogeneous in average stocks, increases in the rate of interest will certainly change the ratios of money stocks in favor of the more expensive stocks.

Thus, if as is sometimes asserted, demand deposits actually have a negative storage cost because of the value of services performed by the bank then increases in the interest rate will increase the currency-deposit ratio.[29] It is interesting to note that the long trend of decreasing currency-deposit ratios has recently reversed itself and that this reversal has coincided with the increasing level of interest rates.[30]

IV. Implications for the Currency-Deposit Ratio

The approach used in this paper can be extended to answer questions concerning the determination of the currency-deposit ratio. In particular, the currency-deposit ratio can be solved for in any of the approaches used by simply limiting the number of money forms to two; currency and demand deposits. In the following, I shall limit my discussion to the approach followed in Section II above.

From the results presented in (12) above the currency-deposit ratio may be written the following way,

$$\frac{m_C}{m_D} = \gamma(\hat{c}, \hat{l}, r, \delta_{\hat{y}}, \delta_{\hat{c}}, \delta_C, \delta_D) \tag{49}$$

where m_C and m_D are real holdings of currency and demand deposits, respectively, δ_C and δ_D are the storage costs of currency and demand deposits, respectively, and the other arguments of (49) are as defined in Section II above.

Now the crucial element in the analysis of the currency-deposit ratio is not (49) but rather the signs of the partial derivatives of (49). These derivatives can be evaluated by first differentiating the quotient m_C/m_D and then assigning the appropriate values to the individual derivatives. First, note that

$$\frac{\partial\left(\frac{m_C}{m_D}\right)}{\partial x_i} = \frac{m_D\frac{\partial m_C}{\partial x_i} - m_C\frac{\partial m_D}{\partial x_i}}{m_D^2} \tag{50}$$

where x_i is the ith argument of (49).

The sign of (50) will depend on the argument chosen and on the assumptions underlying the derivation of the original demand functions. In particular, let me assume that $G_{ij}=0$ for all i and j, $i\neq j$, i.e., the conditions that generated the maximum information concerning the individual derivatives in Section II. In this case, the following solutions for the derivatives of (50) result, and are shown as the system (51). Note that $\alpha=|S|\ m^2_D$, $\beta=(1+G_yH_w)/H_w$ and that G_{CC}, G_{DD} are

[29] See Milton Friedman, (p. 42).

[30] As pointed out in fn. 22 an increase in the rate of inflation is similar in its effect to an increase in the rate of interest. Of course the increase in interest rates referred to is the money rate which in my terms is $(r+\rho)$ where ρ is the rate of inflation.

$$\frac{\partial\gamma}{\partial\hat{c}} = \frac{G_{yy}G_{\bar{v}\bar{v}}G_{\bar{c}\bar{c}}G_c}{\alpha\beta}\left[m_D(r+\delta_C)G_{DD} - m_C(r+\delta_D)G_{CC}\right] \gtreqless 0$$

$$\frac{\partial\gamma}{\partial\hat{l}} = \frac{G_{yy}G_{\bar{v}\bar{v}}G_{\bar{c}\bar{c}}}{\alpha\beta}\left[m_D(r+\delta_C)G_{DD} - m_C(r+\delta_D)G_{CC}\right] \gtreqless 0$$

$$\begin{aligned}\frac{\partial\gamma}{\partial r} = \frac{\beta}{\alpha}\Big\{&(m_CG_{CC} - m_DG_{DD})G_{\bar{v}\bar{v}}G_{\bar{c}\bar{c}} \\ &+ G_{yy}\left[(r+\delta_{\bar{c}})G_{\bar{v}\bar{v}} + (r+\delta_{\bar{v}})G_{\bar{c}\bar{c}}\right]\left[m_C(\delta_{\bar{c}} - \delta_D)G_{CC} - m_D(\delta_{\bar{c}} - \delta_C)G_{DD}\right] \\ &+ (\delta_C - \delta_D)G_{yy}G_{\bar{c}\bar{c}}G_{\bar{v}\bar{v}}\left[m_C(r+\delta_C) + m_D(r+\delta_D)\right]\Big\} \gtreqless 0\end{aligned}$$

(51)
$$\frac{\partial\gamma}{\partial\delta_{\bar{v}}} = \frac{\beta(r+\delta_{\bar{v}})G_{\bar{c}\bar{c}}G_{\bar{v}\bar{v}}}{\alpha}\left[m_D(r+\delta_C)G_{DD} - m_C(r+\delta_D)G_{CC}\right] \gtreqless 0$$

$$\frac{\partial\gamma}{\partial\delta_{\bar{c}}} = \frac{\beta(r+\delta_{\bar{c}})G_{\bar{v}\bar{v}}G_{\bar{y}\bar{y}}}{\alpha}\left[m_D(r+\delta_C)G_{DD} - m_C(r+\delta_D)G_{CC}\right] \gtreqless 0$$

$$\begin{aligned}\frac{\partial\gamma}{\partial\delta_D} = \frac{\beta}{\alpha}\Big\{&\left[m_D(r+\delta_D)(r+\delta_C)G_{yy} + m_CG_{CC}\right]G_{\bar{v}\bar{v}}G_{\bar{c}\bar{c}} \\ &+ m_CG_{yy}\left[(r+\delta_C)^2G_{\bar{v}\bar{v}}G_{\bar{c}\bar{c}} + (r+\delta_C)^2G_{CC}G_{\bar{v}\bar{v}} + (r+\delta_{\bar{v}})^2G_{CC}G_{\bar{c}\bar{c}}\right]\Big\} > 0\end{aligned}$$

$$\begin{aligned}\frac{\partial\gamma}{\partial\delta_C} = -\frac{\beta}{\alpha}\Big\{&\left[m_C(r+\delta_D)(r+\delta_C)G_{yy} + m_DG_{DD}\right]G_{\bar{v}\bar{v}}G_{\bar{c}\bar{c}} \\ &+ m_DG_{yy}\left[(r+\delta_D)^2G_{\bar{v}\bar{v}}G_{\bar{c}\bar{c}} + (r+\delta_{\bar{c}})^2G_{DD}G_{\bar{v}\bar{v}} + (r+\delta_{\bar{v}})^2G_{DD}G_{\bar{c}\bar{c}}\right]\Big\} < 0\end{aligned}$$

---

the second own partial derivatives of the transactions time function with respect to currency and demand deposits, respectively.

Note that even in this extreme case only two of the seven derivatives have determinate signs. These two are the derivatives with respect to the storage costs of the two forms of money. As expected, increases in the currency storage cost decrease the currency-deposit ratio and increases in the demand deposit storage cost increase the currency-deposit ratio. However, if I make the additional assumption that changes in the level of consumption leave the currency-deposit ratio unchanged then the remainder of the derivatives are determinate. In this case the signs of the derivatives in (51) become

(52)
$$\frac{\partial\gamma}{\partial\hat{c}} = \frac{\partial\gamma}{\partial\hat{l}} = \frac{\partial\gamma}{\partial\delta_{\bar{v}}} = \frac{\partial\gamma}{\partial\delta_{\bar{c}}} = 0$$

$$\frac{\partial\gamma}{\partial r} \gtreqless 0 \text{ as } \delta_C \gtreqless \delta_D; \quad \frac{\partial\gamma}{\partial\delta_D} > 0; \quad \frac{\partial\gamma}{\partial\delta_C} < 0$$

Thus, the result of assuming that $\partial\gamma/\partial\hat{c}=0$ is that the currency-deposit ratio depends only on the relative price of currency and demand deposits.

V. Conclusion

In this paper I have been able to demonstrate that the theory of demand for money can be derived from utility maximization without making money holdings an argument of the utility function. Essentially, the approach concentrates on the transaction time saving aspect of using

money. Thus, the results deal with transactions balances only. The demand for money functions derived have all the properties usually assumed in models where the demand for money is a primitive statement in the model.

In the most general case the only derivative of the demand for any particular money form that is determinate is the own storage cost derivative. Thus, the derivative of the demand for a particular money form with respect to the interest rate is indeterminate. This indeterminancy is due to the fact that the relative costs of using the various money forms is dependent on the interest rate. However, even though it cannot be shown in the general case, that the demand for a particular money form is inversely related to interest rates, it can be shown that the demand for the sum of all stocks held is so related.

Since the model allows for an arbitrary number of money forms, it has direct relevance for questions concerning the optimal currency-deposit ratio. In Section IV the implications of the model for the currency-deposit ratio were derived. It was shown that for somewhat restrictive assumptions the currency-deposit ratio derivatives can all be signed. In particular, the currency-deposit ratio is positively related to the costs of holding demand deposits and inversely related to the costs of holding currency. Moreover, if it is assumed that demand deposits are cheaper to hold than currency, as is often asserted, then the currency-deposit ratio is positively related to both the real and money interest rate. This is an interesting result considering the recent rise in both interest rates and the currency-deposit ratio.

REFERENCES

W. J. Baumol, "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quart. J. Econ.*, Nov. 1952, *66*, 545–56.

K. Brunner and A. Meltzer, "Economies of Scale in Cash Balances Reconsidered," *Quart. J. Econ.*, Aug. 1967, *81*, 422–36.

——— **and** ———, "The Uses of Money: Money in the Theory of Exchange," unpublished 1970.

R. Clower, "A Reconsideration of the Micro-Foundations of Monetary Theory," *Western Econ. J.*, Dec. 1967, *5*, 1–8.

M. Friedman, *The Optimum Quantity of Money*, Chicago 1969.

D. Patinkin, *Money, Interest and Prices*, 2d ed., New York 1965.

J. Tobin, "The Interest Elasticity of the Transactions Demand for Money," *Rev. Econ. Statist.*, Aug. 1956, *38*, 241–47.

COMMUNICATIONS

Interest Rates and the Short-Run Consumption Function

By WARREN E. WEBER*

In a previous paper in this *Review*, I used the assumption that a representative consumer maximizes utility over a multiperiod horizon to obtain a consumption function which includes the rate of interest as an independent variable. In that study, interest rates were a statistically significant determinant of annual aggregate consumption. In this note, I examine the question whether interest rate changes are also an important determinant of quarterly consumption movements. I will also determine whether the evidence on consumer behavior obtained with the short-run data is consistent with that obtained from the long-run data.

I. Empirical Analysis

This empirical analysis is performed using the consumption function (10) in my previous study. This consumption function is

$$C_t = g^*(\overset{m}{r}_t, e) \cdot \left\{ W_t + Y_t \sum_{j=0}^{L-e-1} \left[\frac{k_1}{k_3(k_1 + k_2 \overset{m}{r}_t)} \right]^j \right\} + \epsilon_t, \quad (1)$$

where C_t, W_t, and Y_t are aggregate consumption, non-human wealth, and labor income at time t, respectively; r_t^m is a particular market rate of interest at time t; ϵ_t is an independently and identically distributed normal random variable with mean 0 and variance ν^2; β, k_1, k_2, and k_3 are parameters; and $L-e$ is the remaining lifetime of consumers. In (1)

$$g^*(\overset{m}{r}_t, e) = \left[\sum_{j=0}^{L-e-1} (k_1 + k_2 \overset{m}{r}_t)^{-\beta\sigma j} \right]^{-1}$$

The parameter $\sigma = 1/(1+\beta)$ is the partial intertemporal elasticity of substitution for consumption. The economic meaning of the other three parameters in the model is best understood if we let γ be consumers' subjective rate of discount for future consumption, ξ be consumers' anticipated rate of growth of labor income, and θ_1 and θ_2 be two parameters which relate the rate of interest at which consumers expect to borrow and lend in all periods (r_t) and a particular market interest rate according to

$$r_t = \theta_1 + \theta_2 \overset{m}{r}_t \quad (2)$$

Then $k_1 = (1+\gamma)^{-1/\beta}(1+\theta_1)$, $k_2 = (1+\gamma)^{-1/\beta}\theta_2$, and $k_3 = (1+\theta_1)/(1+\xi)$.

The results of the empirical analysis using aggregate quarterly data for the United States for the period 1952–62 are presented in two tables below.[1] In Table 1 we present the results when consumers are assumed to base their interest rate expectations upon a single market interest rate. Five different interest rates are used in the analysis to determine whether the choice of interest rate affects the empirical results.[2]

* Assistant professor of economics, Virginia Polytechnic Institute.

[1] The observation period is limited to 1952–62 because the Federal Reserve Board's Flow-of-Funds data used in the construction of both the aggregate non-human wealth and aggregate consumption series are not available for more recent quarters. The data series and description of the method of construction will be sent upon request. All flow data is expressed in terms of quarterly rates and in real terms. I set the remaining lifetime of consumers ($L-e$) equal to 120 quarters. This is the lifetime used in the previous paper.

[2] The five interest rate series are Moody's Aaa and Baa corporate bond yields, the market yields on 3-month bills and 3-to-5 year *U.S.* government taxable securities, and the yield on long-term *U.S.* government bonds. All interest rate series are obtained from the *Supplement to Banking and Monetary Statistics*, Section

TABLE 1—CONSUMPTION FUNCTION ESTIMATES

Parameter	Moody's		U.S. Government Securities		
	Aaa	Baa	3-month bills	3- to 5-year issues	long-term
β	4.573	2.849		4.465	6.541
k_1	1.001	0.9917		1.019	0.9953
k_2	0.1739	0.3749	0.000	0.0146	0.2235
k_3	1.005	0.9949		1.028	0.9980
Sum of sq. residuals	35.61	32.00	45.12	44.75	38.22
Test Stat.	10.4	14.8	0.000	0.362	7.30
Sig. Level	0.001	0.001	0.000	0.7	0.01

When $\theta_2 = k_2 = 0$, changes in the rate of interest will have no effect on short-run consumption decisions. However, if $k_2 > 0 (\theta_2 > 0)$, there is an interest rate effect. Shown in the next to last line of Table 1 are the likelihood ratio test statistics of the null hypothesis that $k_2 = 0$ against the alternative that $k_2 > 0$. In the last line of the table we present the significance levels at which the null hypothesis can be rejected.[3] For both of the Moody's corporate bond rates we can reject the null hypothesis at the 0.001 level of significance. For the long-term government bond rate, we can reject the null hypothesis at the 0.01 level of significance. We cannot reject the null hypothesis for reasonable significance levels for either of the short-term government rates.[4]

Thus, on the basis of our findings in Table 1, we can say that the choice of interest rate does have an important effect on the explanatory power of the model. Changes in corporate bond rates do have effects on short-run consumption, whereas changes in short-term government rates do not. We do not have enough information to make a statement either way about long-term government bond rates.

Now suppose that the explanatory power of an interest rate series is directly related to its reliability as an indicator of the interest rates facing consumers. Then the results indicate that corporate bond rates are more reliable indicators than government bond rates. They also indicate that corporate bond rates are more reliable the higher the default risk of the bond and that government bonds rates are more reliable the longer the maturity of the issue.

We now proceed to allow two interest rates to affect consumers' interest rate expectations by changing equation (2) to

$$r_t = \theta_1 + \theta_2 r_t^m + \theta_3 x_t, \tag{3}$$

where x_t is a second market rate of interest. Using (3) and performing the same manipulations which I used to obtain (10) in my previous study, I obtain a new consumption function which differs from (1) in that wherever $(k_1 + k_2 r_t^m)$ appears in (1), it is now replaced by $(k_1 + k_2 r_t^m + k_4 x_t)$, where $k_4 = (1+\gamma)^{-1/\beta}\theta_3$.

Table 2 presents the results when two interest rates are allowed to affect consumers' interest rate expectations. *Ceteris paribus*, an increase in the second market interest rate will decrease consumers' interest rate expectations. However, since $|\theta_2/\theta_3| = |k_2/k_4|$, θ_3 is always less in absolute value than θ_2, so that an equal increase in both market interest rates will cause consumers' expected interest

12. We will use the term "short-term government rates" to refer to the yields on 3-month bills and 3-to-5 year issues. The interest rate data are available upon request.

[3] The level of significance at which the null hypothesis can be rejected is obtained by comparing the value of the test statistic against the percentiles of the chi-square distribution with one degree of freedom. The method of calculation of the test statistics is given in Weber, page 8 and fn. 14. The sum of the squared residuals under the null hypothesis is 45.12.

[4] For 3-month bills the minimum sum of the squared residuals which we obtained is the same as that under the null hypothesis. For this case we cannot empirically distinguish β, k_1, and k_3. Hence, no values for these parameters are reported in the table.

TABLE 2—CONSUMPTION FUNCTION ESTIMATES WITH TWO INTEREST RATES

r_t^m	Moody's Baa				U.S. Gov't 3–5 Year	U.S. Gov't Long-Term	
x_t	Moody's	U.S. Government Securities			U.S. Government		
Parameter	Aaa	3-Month Bills	3- to 5-Year Issues	Long-Term	3-Month	3-Month Bills	3- to 5-Year Issues
β	2.800	2.957	2.597	2.880	4.249	14.95	8.769
k_1	0.9917	0.9919	0.9917	0.9917	1.008	0.9963	1.009
k_2	0.9621	0.4938	0.6800	0.8103	0.2148	0.2679	0.2186
k_3	0.9943	0.9947	0.9943	0.9944	1.013	0.9981	1.015
k_4	−0.6828	−0.2234	−0.3579	−0.5629	−0.1811	−0.1075	−0.1066
Sum of sq. Residuals	28.92	24.44	21.75	28.13	38.12	29.54	28.37
Ω_1	4.73	12.1	17.3	5.95	7.06	11.3	13.1
Sig. Level	0.05	0.001	0.001	0.025	0.01	0.001	0.001
Ω_2	19.6	27.0	32.1	21.5	7.42	18.6	20.4
Sig. Level	0.001	0.001	0.001	0.001	0.025	0.001	0.001

rate to increase. Also note that when two corporate bond rates are used in the analysis, it is the one with the lower default risk which enters with the negative sign and has the smaller effect. When a corporate rate and a government rate are used, it is the government rate which has the negative sign and the smaller effect. When two government rates are used, it is the one with the shorter maturity which has the negative sign and the smaller coefficient. In all cases, it is the market interest rate which had the higher explanatory power alone which has the larger coefficient when combined with a second interest rate. Thus, these results tend to reinforce my conclusions about the relative reliability of the various interest rates as indicators of the interest rate conditions facing consumers.

When two interest rates affect consumers' expectations, changes in market interest rates will not affect short-run consumption only if $\theta_2=\theta_3=0$; i.e., only if $k_2=k_4=0$. There will be an interest rate effect if either θ_2 or θ_3 is not equal to zero; i.e., if either $k_2\neq0$ or $k_4\neq0$. The line labelled Ω_2 in Table 2 presents the likelihood ratio test statistics of the null hypothesis that $k_2=k_4=0$ against the alternative hypothesis that either $k_2\neq0$ or $k_4\neq0$. The next line shows the levels of significance at which the null hypothesis can be rejected. With the exception of the case in which two short-term government rates are used in the analysis, we can reject the null hypothesis of no interest rate effect at the 0.001 level of significance. These results are much stronger evidence in favor of the contention that interest rates are an important determinant of short-run consumption than are those presented in Table 1.

Does the inclusion of the second interest rate in the analysis significantly increase the explanatory power of the model? The line labelled Ω_1 in Table 2 presents the likelihood ratio test statistics of the null hypothesis that $k_4=0$ $(\theta_3=0)$ against the alternative hypothesis that $k_4\neq0$ $(\theta_3\neq0)$. The next line presents the levels of significance at which the null hypothesis can be rejected. With Moody's Baa rate we obtain a significant (at the 0.001 level) increase by including either the yield on 3-month bills or 3-to-5 year issues. However, the null hypothesis is not rejected when either the Aaa rate or the long-term government bond rate is included with the Baa rate. Including either the yield on 3-month bills or 3-to-5 year issues with the long-term government rate significantly increases the model's explanatory power. These results coupled with those presented directly above indicate that two market interest rates provide a better explanation of short-run consumption than does a single market interest rate. These results also lead

TABLE 3—COMPARISON OF ESTIMATES USING ANNUAL AND QUARTERLY DATA

Parameter	Annual Data: Durand's Basic Yields Years to Maturity 1	5	30	Moody's Baa	Av. Annual Yield Time & Saving Deposits In Com. Banks	Quarterly Data: Moody's Aaa	Moody's Baa	U.S. Gov't Securities 3–5 Year	Long-Term
β	6.288	2.599	1.427	6.675	2.778	4.573	2.849	4.465	6.541
$\sigma=1/(1+\beta)$	0.1372	0.2779	0.4120	0.1303	0.2647	0.1794	0.2598	0.1830	0.1326
γ	−0.3140	−0.1564	−0.0820	0.0274	−0.1446	0.0174	0.0388	−0.0593	0.0658
ξ	−0.1092	−0.0949	−0.0668	−0.0162	−0.083	0.0003	0.0108	−0.0225	0.0070
θ_2	0.1839	0.3500	0.9248	0.5928	0.4295	0.0437	0.0950	0.0036	0.0564

us to conclude that a combination of the Moody's Baa rate and a short-term government rate provides the best indication of the interest rates facing consumers. The expected interest rate calculated from a combination of yields on two different types of bonds provides a more reliable interest rate indicator than either a combination of two corporate bond rates or two government security rates.[5]

II. Comparison of the Long-Run and Short-Run Results

Table 3 presents the estimates of the parameters β, θ_2, γ, and ξ with both annual and quarterly data.[6] The quarterly estimates are obtained from Table 1. The annual estimates are obtained from Tables 1 and 2 in my *Review* article.

Our estimates of the partial intertemporal elasticity of substitution (σ) are similar and uniformly less than 0.5 for both sets of data. These results indicate that consumers do not regard consumption in different quarters to be any better substitutes than they regard consumption in different years. And, regardless of the time period over which the flow of consumption is measured, they regard consumption in different time periods to be poor substitutes.

The two sets of data yield conflicting results concerning the consumers' rate of time preference (γ). The annual data indicate that consumers are impatient, desiring to advance the timing of their consumption. The quarterly data indicate that they are slightly negatively impatient, desiring to slightly postpone the timing of their satisfaction. This difference could arise because next quarter's consumption appears relatively near to this quarter's, so that consumers are almost indifferent whether they have their consumption in this period or the next. However, next year's consumption is relatively far off, so that they may want some of next year's consumption now.

The two sets of data also yield conflicting results about consumers' anticipated rate of growth of labor income (ξ). The annual data indicate that consumers expect labor income to decline over the remainder of their economic lifetimes, whereas the quarterly data indicate that they expect it to increase. This conflict may only be illusory, however, for when we took account of retirement in the study with annual data we found that our estimates of ξ became positive. Since

[5] I also performed the analysis using the rate of inflation rather than a second market interest rate for x_t. The rationale for including the rate of inflation is that consumers may base their consumption decisions upon real rather than nominal interest rates. Consequently, I expected and found negative signs on k_4. I also found however, that including the rate of inflation only trivially increased the explanatory power of our model, and thus concluded that it was nominal interest rate changes which affected short-run consumption decisions.

[6] We cannot obtain direct estimates of θ_1, θ_2, γ, and ξ unless we are willing to arbitrarily specify a value for one of these parameters. This same problem arose in my previous study; the solution was to note that θ_1 could be interpreted as an interest rate floor or a risk premium which gave a basis upon which to assign a value to this parameter. I chose the value $\theta_1=0.02$ since it was approximately the mean of the average annual yield on time and savings deposits in commercial banks over the period 1930–1965. I set $\theta_1=0.005$ here in order to make the estimates with quarterly data comparable with the estimates from the annual data.

retirement is relatively further off for quarterly consumption decisions, it may not exert the influence on ξ which it apparently does for the annual decisions, so that ξ becomes positive with the quarterly data.

Finally, we find that our estimates of θ_2 are uniformly smaller with the quarterly data than with the annual data. This result is certainly expected and consistent. In the annual study the average effective rate of interest is an annual rate, whereas in the quarterly study it is a quarterly rate. Since a 1 percent increase in a quarterly rate is approximately equal to a 4 percent increase in an annual rate, we would expect changes in the market rate (always an annual rate) to have a smaller effect on consumers' average effective interest rate in the quarterly study.[7] In addition, the two sets of data yield consistent results concerning the magnitude of θ_2 and the type of asset yield upon which consumers are basing their interest rate expectations. We find that θ_2 is larger the longer the maturity of the asset and the greater the default risk of the asset.

III. Conclusion

In my previous study with annual data I found interest rates to be an important determinant of consumption, regardless of the interest rate series used in the empirical analysis. The results obtained with the quarterly data do not allow such a strong conclusion. Here we find that when a single market interest rate is used in the analysis, the choice of interest rate makes a difference in the empirical results. Only corporate bond rates are found to be statistically significant determinants of consumption. It is only when two market interest rates are included in the analysis that we obtain results which are in any way as insensitive to the choice of interest rate series as were the annual results. And even in this case, we find that using only two short-term government rates as explanatory variables does not improve the explanatory power of the model.

All of the evidence presented in this and the previous study supports the hypothesis that interest rates influence consumption. Since all of the empirical evidence in this study which does not support this hypothesis is obtained using exclusively yields on government securities, I feel that the evidence presented for the hypothesis outweighs that presented against it.

[7] As this discussion clearly shows, in the annual study θ_2 is a unitless number. However, in the quarterly study it is in terms of years/quarter. Therefore, each of the quarterly indirect estimates of θ_2 are divided by four to get the results presented in Table 3.

REFERENCES

W. E. Weber, "The Effect of Interest Rates on Aggregate Consumption," *Amer. Econ. Rev.*, Sept. 1970, *60*, 591–600.

Board of Governors of the Federal Reserve System, "Money Rates and Securities Markets," *Supplement to Banking and Monetary Statistics, Section 12*, Washington 1966.

Unemployment and Inflation: A Cross-Country Analysis of the Phillips Curve

By David J. Smyth*

This paper undertakes a cross-section analysis of the relation between unemployment and the rate of inflation using data for eleven countries for the period 1950–60. A markedly non-linear Phillips curve is obtained.[1] The paper also investigates two further hypotheses. First, that a single Phillips curve should be replaced by a family of curves, one for each rate of productivity increase. Secondly, that the average rate of inflation will be higher for a given average level of unemployment, the greater the magnitude of cyclical variations in unemployment. Both these hypotheses are rejected.

The plan of the paper is as follows. Section I describes the data used and the notation. Section II estimates the cross-country Phillips curve. The hypotheses relating to productivity growth and cyclical variations in unemployment are investigated in Sections III and IV, respectively. The paper's conclusions are summarized in Section V.

I

A major difficulty in any study involving comparison of unemployment rates in different countries is that the coverage of unemployment statistics varies widely from country to country. It is important to attempt to convert the estimates to a common basis. Angus Maddison assumes that the census coverage of unemployment is fairly comparable between countries and uses information for census years to make corresponding adjustments for other years.[2] His estimates have been adopted in the present paper.

Eleven countries are used in this study: they are Belgium, Canada, Denmark, France, Germany (Federal Republic), Italy, Netherlands, Norway, Sweden, United Kingdom, and the United States. The availability of Maddison's comparable unemployment series limits the sample to this size;[3] it also dictates the choice of period studied, 1950–60.

The average annual increase in the price index relating to the gross national product over the period 1950–60 is used as the measure of the rate of inflation. (See Maddison, p. 45.) Three measures of productivity growth are used: the average annual rate of growth of output per man-hour, of output per head of population, and of output (see Maddison, pp. 28, 30, and 37). The variance and the coefficient of variation of unemployment over the eleven years 1950–60 are used as measures of the time dispersion of unemployment. These are both calculated from annual data.

The following notation is used.

P = average annual increase in the gross national product price index, 1950–60.

* Professor of economics, Claremont Graduate School.
I have benefited from discussion with Arthur Butler and James Holmes.

[1] The Phillips curve has been widely interpreted as providing a relationship between the rate of inflation and unemployment (see, for instance, Paul Samuelson and Robert Solow) although in his original article Phillips used change in money wage rates and not change in prices.

[2] Maddison also makes some other adjustments. Details of his procedures and his unemployment series are given in his Appendix E.

[3] Maddison also gives unemployment estimates for Switzerland. However, this country was eliminated from the analysis because of difficulties in reconciling the data used by Maddison and that given in the United Nations *Statistical Yearbook*. The census estimate of the unemployment rate in Switzerland on December 1st, 1950, was 0.42 percent while the registered unemployment series gave a figure of 0.29 percent for November 30th of the same year. (Both figures are from the *Annuaire Statistique de la Suisse*, 1958, and are reported by Maddison (p. 222).) These estimates seem extraordinarily low. The figure for 1950 given in the *United Nations Statistical Yearbook* is 1.8 percent. In view of this wide discrepancy it seemed best to exclude Switzerland from the sample.

U = average unemployment rate, 1950–60.
G_1 = average annual rate of growth of output per man-hour, 1950–60.
G_2 = average annual rate of growth of output per head of population, 1950–60.
G_3 = average annual rate of growth of output, 1950–60.
V = variance of annual unemployment rates, 1950–60.
C = coefficient of variation of unemployment rates, $(= V^{1/2}/U)$, 1950–60.

II

Figure 1 presents the scatter diagram between the average annual price increase and the average unemployment rate. It is apparent that the relationship between P and U is a highly non-linear one. Accordingly, we follow Richard Lipsey and George Perry and make use of non-linear transformations of unemployment.[4] The most satisfactory regression equation obtained is

$$(1) \quad P = 2.73 + 9.48U^{-2} \qquad R^2 = .869$$
$$(0.29) \quad (1.22)$$

The result obtained when U^{-1} is used instead of U^{-2} is inferior and there is little point in using U^{-1} as well as U^{-2} as the two are highly correlated.[5] If U is used instead of U^{-2} we have

$$(2) \quad P = 6.70 - 0.655U \qquad R^2 = .548$$
$$(0.81) \quad (0.198)$$

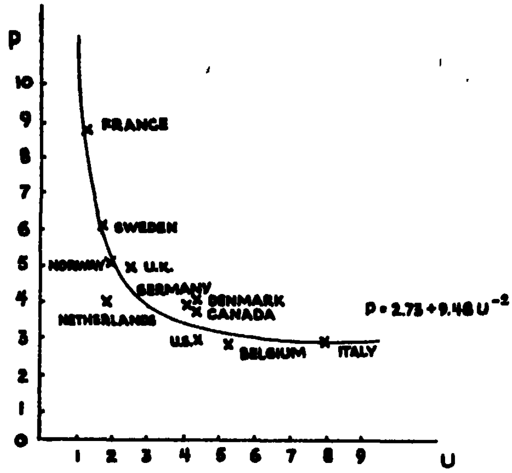


FIGURE 1. AVERAGE RATE OF INFLATION AND AVERAGE UNEMPLOYMENT RATE, 11 COUNTRIES, 1950–60.

This linear relation is markedly inferior to that involving U^{-2}; the F-ratio between the two mean square deviations is 3.46 which is significant at the 5 percent level.

The fit obtained in (1) between the rate of inflation and the unemployment rate is remarkably good for a cross-section analysis.[6] The coefficient of U^{-2} is easily significant at the 0.1 percent level. The regression line given by (1) is plotted in Figure 1. As the unemployment rate falls below about 2 percent the rate of inflation increases markedly and as unemployment rises above about 4 percent little diminution in the rate of inflation results. Given that governments have a trade off between the rate of inflation and the unemployment rate, this suggests that the optimum rate of unemployment is likely to lie between 2 and 4 percent unless one of the policy objectives is weighted relatively heavily compared with the other.[7]

III

Perry has advocated the replacement of a single Phillips curve by a family of curves,

[4] Adoption of Lipsey's mathematical form involves no commitment to his derivation of the Phillips curve. James Holmes and I have demonstrated that there is not a unique relationship between the excess demand or supply of labor and unemployment and that consequently the Phillips curve cannot be obtained from Lipsey's model.

[5] Standard errors are given in brackets. When U^{-1} is used instead of U^{-2} the regression obtained is

$$P = 1.46 + 7.89U^{-1} \qquad R^2 = .810$$
$$(0.52) \quad (1.27)$$

The correlation between U^{-1} and U^{-2} is .981. When U^{-1} and U^{-2} are included in the same regression this multicollinearity is reflected in increased standard errors, the estimated regression being

$$P = 3.32 - 3.45U^{-1} + 13.4U^{-2} \qquad R^2 = .875$$
$$(1.01) \quad (5.64) \quad (6.5)$$

[6] It is noteworthy that the biggest deviation from the regression line is for the Netherlands which is substantially negative presumably reflecting the success of its well-known incomes policy during the period.

[7] Converted into terms of the regular $U.S.$ unemployment series this range becomes 2.4 to 4.7 percent.

each curve corresponding to a different rate of productivity increase. According to Perry's hypothesis, a rise in the rate of productivity increase will shift the Phillips curve towards the origin; correspondingly a fall in the rate of productivity increase will shift the Phillips curve out away from the origin. On the other hand, Paul Streeten believes that productivity increases are accompanied by equivalent rises in factor payments resulting in no diminution in inflationary pressures.

The present section investigates whether a country's rate of inflation is related to its rate of productivity increase. Three measures of productivity growth are used: G_1, the rate of growth of output per man-hour; G_2, the rate of growth of output per head of population; and G_3, the rate of growth of output. The following regressions are obtained when these three measures of productivity growth are in turn included with U^{-2}.

(3) $P=2.70+9.47U^{-2}+0.014G_1 \quad R^2=.870$
$\quad (0.72) \; (1.32) \quad (0.201)$

(4) $P=2.70+9.48U^{-2}+0.007G_2 \quad R^2=.870$
$\quad (0.54) \; (1.30) \quad (0.142)$

(5) $P=2.76+9.48U^{-2}-0.009G_3 \quad R^2=.870$
$\quad (0.71) \; (1.30) \quad (0.150)$

It is clear from equations (3) to (5) that the inclusion of the productivity variables adds nothing to the explanation of the rate of inflation. A country's rate of inflation appears to be independent of its rate of productivity increase. The cross-country evidence thus supports Streeten's position rather than Perry's.

IV

Lipsey has suggested that if the Phillips curve is non-linear, then the position of the observed Phillips curve will vary with the dispersion of unemployment. Lipsey's argument may be summarized as follows. Suppose that there is more than one labor market, each market having its own identical convex Phillips curve. If unemployment is not the same in each market, then the fitted aggregate curve will lie above the individual market curves. An increase in the dispersion of unemployment will raise the aggregate curve, a decrease will lower it. Using geographical and industrial subaggregates G. C. Archibald and Anthony Thirwall have recently provided support for Lipsey's hypothesis.

We expect to observe the same sort of phenomena as unemployment fluctuates over time. If a country's unemployment rate is not constant but varies cyclically, then the point relating its average price change to its average unemployment rate should lie above the true Phillips curve. Further, for a given average unemployment rate, the point should be higher the greater the dispersion of unemployment about its average over the period under consideration: thus countries that have experienced large cyclical fluctuations during the period 1950–60 should have higher rates of inflation than countries experiencing lesser cyclical fluctuations. To test this hypothesis we calculate two measures of the time dispersion of unemployment from the annual observations for 1950–60, the variance, V, and the coefficient of variation, C, of each country's unemployment series.

Equations (6) and (7) report the effect of including V and C with U^{-2}.

(6) $P=2.51+10.00U^{-2}+0.099V \quad R^2=.873$
$\quad (0.56) \; (1.72) \quad (0.221)$

(7) $P=3.32+8.88U^{-2}-1.99C \quad R^2=.884$
$\quad (0.66) \; (1.37) \quad (2.00)$

In neither regression is the dispersion variable significant at the 10 percent level and in (7), it has the wrong sign. We thus conclude that a country's rate of inflation during the period 1950–60 is not influenced by the magnitude of its cyclical fluctuations during the same period and that no support is provided for Lipsey's dispersion of unemployment hypothesis.[8]

[8] Various combinations of U, U^{-1}, U^{-2}, G_1, G_2, G_3, V, and C, other than those reported in this paper, were also investigated. In no regression were any of the productivity or dispersion variables significant at the 10 percent level.

V

The conclusions of this paper may be simply stated.

A cross-country analysis of the Phillips curve for the period 1950–60 yielded a downward sloping curve; this curve was markedly convex towards the origin. For a cross-section analysis the fit was remarkably good.

The hypothesis that a single Phillips curve should be replaced by a family of curves, each curve representing a different rate of productivity increase, was tested and rejected; a country's rate of productivity increase had no effect upon its rate of inflation.

The hypothesis that the time dispersion of unemployment influenced a country's rate of inflation was examined. The dispersion variables were not significant, a country's inflation rate thus being independent of the extent of cyclical fluctuations in unemployment.

REFERENCES

G. C. Archibald, "The Phillips Curve and the Distribution of Unemployment," *Amer. Econ. Rev.*, May 1969, *59*, 124–34.

J. M. Holmes and D. J. Smyth, "The Relation Between Unemployment and Excess Demand for Labor: An Examination of the Theory of the Phillips Curve," *Economica*, Aug. 1970, *37*, 311–15.

R. G. Lipsey, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1862–1957: A further Analysis," *Economica*, Feb. 1960, *27*, 1–31.

A. Maddison, *Economic Growth in the West*, New York 1964.

G. L. Perry, *Unemployment, Money Wage Rates, and Inflation*, Cambridge, Mass. 1966.

A. W. Phillips, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957," *Economica*, Nov. 1958, *25*, 283–99.

P. A. Samuelson and R. M. Solow, "Analytical Aspects of Anti-Inflation Policy," *Amer. Econ. Rev. Proc.*, May 1960, *50*, 177–94.

P. Streeten, "Productivity Inflation," *Kyklos*, 1962, *15*, 723–31.

A. P. Thirwall, "Demand Disequilibrium in the Labour Market and Wage Rate Inflation in the United Kingdom," *Yorkshire Bull. Econ. Soc. Res.*, May 1969, *21*, 66–76.

United Nations, *Statistical Yearbook*, New York 1957.

Long-Run Scale Adjustments of a Perfectly Competitive Firm and Industry: An Alternative Approach

By RICHARD D. PORTES*

The purpose of this note is to use the cost function technique to derive the results of the recent article in this *Review* by C. E. Ferguson and Thomas Saving (hereafter F-S). Our interest lies not so much in the results as in the simplicity and power of the technique itself, in contrast to the rather cumbersome conventional apparatus of comparative statics micro-theory.[1] We shall not, therefore, discuss in detail the F-S interpretations of their results, although the technique should give the reader a more intuitive grasp of what is going on.

Following F-S, we assume that the production function $f(x)$ is increasing in each input, strictly concave and twice continuously differentiable on the positive orthant. All equilibrium input vectors are assumed strictly positive. We start from the duality between production and cost functions demonstrated by Shephard, Uzawa, and McFadden; under our assumptions, the production function may be uniquely characterized by its minimum total cost function, defined for all $p>0$ by[2]

$$C = \min_{x \geq 0} \left\{ \sum p_i x_i \,\middle|\, f(x) = q \right\} = C(p_1, \ldots, p_n; q) \tag{1}$$

The cost function is nondecreasing (increasing at an equilibrium point) and strictly concave in input prices p, and it is increasing and strictly convex in output q. It is twice continuously differentiable in all variables.

We must emphasize here that the duality relationship means that we can equally well begin by assuming these properties of the cost function and omitting any explicit consideration of the production function. A theory of production can take as its "primitive concept" the production function, the technology set, or the cost or profit function. There is now a well-developed theory of the relationships between these concepts (see Shephard (1970)). In any given context, the choice is therefore a matter of suitability to the problem at hand. We would suggest that the cost function is most convenient for a wide variety of problems in the theory of production and demand, from relatively simple questions, like the present one, to much more complex investigations (e.g., Gorman).

We shall write $C' = \partial C/\partial q$, $C_i = \partial C/\partial p_i$, and $C_{ij} = \partial^2 C/\partial p_i \partial p_j$; and we shall use dC/dp_i when we wish to denote that output is being allowed to vary (while other input prices remain constant).

It is a basic result from cost function theory (see Shephard, McKenzie) that under our assumptions,

$$C_i(p_1, \ldots, p_n; q) = x_i(p_1, \ldots, p_n; q), \quad i = 1, \ldots, n, \tag{2}$$

where $x(p; q)$ is the cost-minimizing input vector for q.[3] We shall write $\partial x_i/\partial p_j = C_{ij}$

* Assistant professor of economics and international affairs at Princeton University. I wish to acknowledge helpful comments from John Black and the referee.

[1] There is at present no single, general, readily available reference covering the theory of cost and profit functions. This is a pity, since the approach would be quite at home in textbooks on intermediate micro-theory. The original reference is R. W. Shephard (1953). The theory was developed further by Hirofumi Uzawa and surveyed by Daniel McFadden. A forthcoming book by Shephard will give by far the most complete development. Applications can be found in Lionel McKenzie, W. M. Gorman (who uses the technique to deal with a problem in the theory of aggregation and gives a detailed, rigorous treatment of profit functions), and Portes (1969).

[2] Our assumptions ensure that the minimum is in fact attained, so that we can write "min" instead of "inf." We use upper-case C to denote the cost *function*, while F-S use lower-case c to denote total cost.

[3] Note that this is F-S (1.8), the essential result from their Appendix Note 1, which underlies the qualitative conclusions of their Section II.

when output is held constant and dx_i/dp_j when output varies.

Since $C(\ )$ is strictly concave in p and strictly convex in q, we have

$$(3)\qquad \frac{\partial x_i}{\partial p_i} = C_{ii} < 0, \qquad i = 1, \ldots, n; \qquad C'' > 0$$

Since the second-order derivatives of $C(\ )$ are continuous, we can reverse the order of differentiation. Noting that C' = marginal cost, we have

$$(4)\qquad \frac{\partial}{\partial p_i}(MC) = \frac{\partial C'}{\partial p_i} = C_i' = \frac{\partial x_i}{\partial q}, \qquad i = 1, \ldots, n$$

$$(5)\qquad \frac{\partial x_j}{\partial p_i} = C_{ji} = C_{ij} = \frac{\partial x_i}{\partial p_j}, \qquad i, j = 1, \ldots, n$$

The ith input is inferior for given p, over some range of q, if $\partial x_i/\partial q < 0$.[4] We note from (4) that MC at given q varies inversely or directly with p_i according as x_i is inferior or not, i.e., an increase in the price of an inferior input will shift MC downward over the range of outputs for which the input is inferior. F-S obtain this conclusion from their (1.5), which has the same qualitative content as our (4).

All this is unavoidable background, and we have already covered some of F-S. We can now obtain the remaining F-S propositions very quickly indeed. Long-run[5] competitive equilibrium for both firm and industry requires MC = price = AC, i.e., $C' = C/q$. Differentiate with respect to p_i, allowing output to vary:

$$\frac{dC'}{dp_i} = C_i' + C''\frac{dq}{dp_i} = \frac{\partial x_i}{\partial q} + C''\frac{dq}{dp_i},$$

using (4); and

$$\frac{d}{dp_i}\left(\frac{C}{q}\right) = \frac{1}{q^2}\left[q\left(C_i + C'\frac{dq}{dp_i}\right) - C\frac{dq}{dp_i}\right] = \frac{x_i}{q},$$

using (2) and the assumption that $C' = C/q$. Equating these expressions and rearranging, we have

$$(6)\qquad \frac{dq}{dp_i} = \frac{x_i}{qC''}\left(1 - \frac{q}{x_i}\cdot\frac{\partial x_i}{\partial q}\right)$$

It is easy to show that this is identical to F-S (21), the key result of Section I of F-S (see the Appendix to this paper, where I show that their η_{ic} equals my $q/x_i \cdot \partial x_i/\partial q$ and their F/cF^* equals my $1/qC''$). Precisely the same qualitative conclusions follow from it: since $C''>0$, we have $dq/dp_i \gtrless 0$ according as $\eta_{ic} \lessgtr 1$.

In deriving (6), we found $d(AC)/dp_i = x_i/q$; since by hypothesis $C' = C/q$, we have $d(AC)/dp_i = x_i C'/C$, which is F-S (3.3), the basis of their Appendix Note 3 (p. 783).

Our approach now allows us to give both more precision and more intuitive content to the diagrammatic interpretation in F-S Section III. An input price change shifts the MC curve and the AC curve. In F-S (17) and (21), as in my (6), we can recognize a decomposition of the response of output into a part associated with the shift of the MC curve with constant output price and a part associated with the move along the MC curve caused by the increase of minimum $LRAC$, which determines output price in the long run. We can see this in the following way. Let output price be r. Then we have

$$(7)\qquad \frac{dq}{dp_i} = \left(\frac{\partial q}{\partial p_i}\right)_{dr=0} + \frac{\partial q}{\partial r}\frac{\partial r}{\partial p_i}$$

Now $\partial q/\partial p_i$ is just the ***short-run*** response of equilibrium output to the change in input price; the firm adjusts to the shift of the MC curve, taking output price as constant. In

[4] For a brief discussion of input inferiority and further references, see Portes (1968).

[5] Note that the distinction between short and long run in F-S, which we follow, is not in terms of whether there are fixed inputs or not—throughout, all inputs are allowed to vary—but rather of whether output price is fixed or allowed to change as the industry adjusts to cost changes.

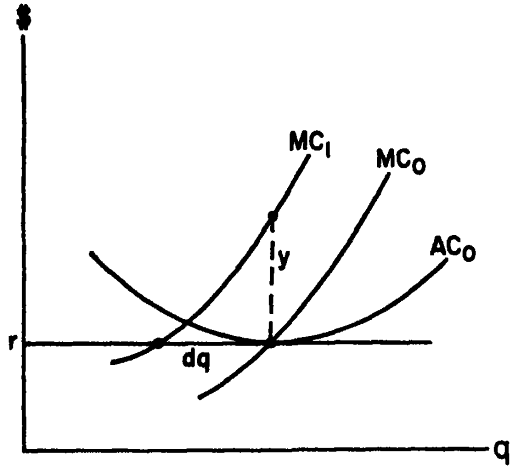


FIGURE 1

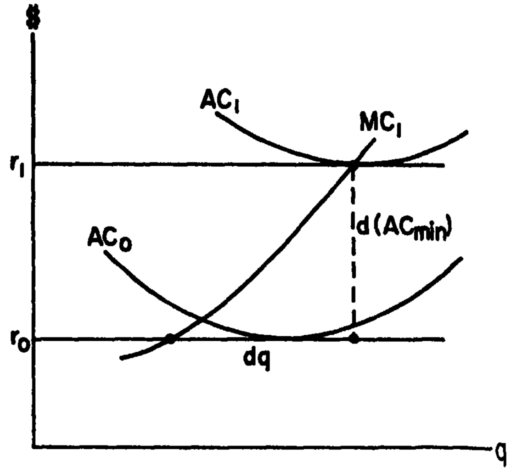


FIGURE 2

the long run, however, output price must change, since the minimum *LRAC* of all firms will have risen due to the input price increase. Consider the second term on the right-hand side of (7): $\partial q/\partial r$ is the change in q per unit movement upward *along* the *MC* curve, the movement upward being measured along the vertical axis; $\partial r/\partial p_i$ is the amount of upward movement per unit increase in p_i, i.e., it is the distance by which minimum *AC* shifts upward as a result of a unit increase in p_i. Thus we can write

$$\frac{\partial q}{\partial r} = \frac{\partial q}{\partial(MC)} = \frac{\partial q}{\partial C'} = \frac{1}{\partial C'/\partial q} = \frac{1}{C''}$$

$$\frac{\partial r}{\partial p_i} = \frac{\partial(AC)}{\partial p_i} = \frac{\partial(C/q)}{\partial p_i} = \frac{C_i}{q} = \frac{x_i}{q},$$

so that

$$(8) \qquad \frac{\partial q}{\partial r}\frac{\partial r}{\partial p_i} = \frac{x_i}{qC''}$$

We then have

$$(9) \qquad \frac{dq}{dp_i} = \left(\frac{\partial q}{\partial p_i}\right)_{dr=0} + \frac{x_i}{qC''}$$

This is equivalent to the decomposition in F-S (17). Indeed, (7) and (9) are just a different way of writing (6), and comparing (9) with (6) gives us the information that

$$(10) \qquad \left(\frac{\partial q}{\partial p_i}\right)_{dr=0} = -\frac{1}{C''}\cdot\frac{\partial x_i}{\partial q}$$

A simple geometrical interpretation of (10) and (8) is shown in Figures 1 and 2.

In Figure 1, we show $(\partial q/\partial p_i)_{dr=0}$, the change due to the shift of the *MC* curve. From the geometry, $y = -C''dq$, so $dq = -y/C''$. But y is the vertical shift of the *MC* curve due to the change dp_i, so

$$y = (\partial C'/\partial p_i)dp_i = C_i' dp_i = (\partial x_i/\partial q)dp_i$$

Thus $dq = -(1/C'')(\partial x_i/\partial q)dp_i$, which gives (10).

In Figure 2, we show $(\partial q/\partial r)(\partial r/\partial p_i)$, the change due to the move along the *MC* curve. As we found when deriving (6), $d(AC) = (x_i/q)dp_i$. From the geometry, $d(AC) = C''dq$. Thus $dq = d(AC)/C'' = (x_i/qC'')dp_i$, which gives (8).[6]

Combining the two figures, we get a somewhat more elaborate version of F-S Figure 1 which is our Figure 3. In each panel, the move from *R* to *S* is the first term on the right-hand side of (9), as given in (10) and my Figure 1, while the move from *S* to *T* (measured horizontally) is the second term

[6] Here $d(AC)$ is the change in *minimum* average cost. The result $d/dp_i(C/q) = x_i/q$ used the assumption that $C' = C/q$.

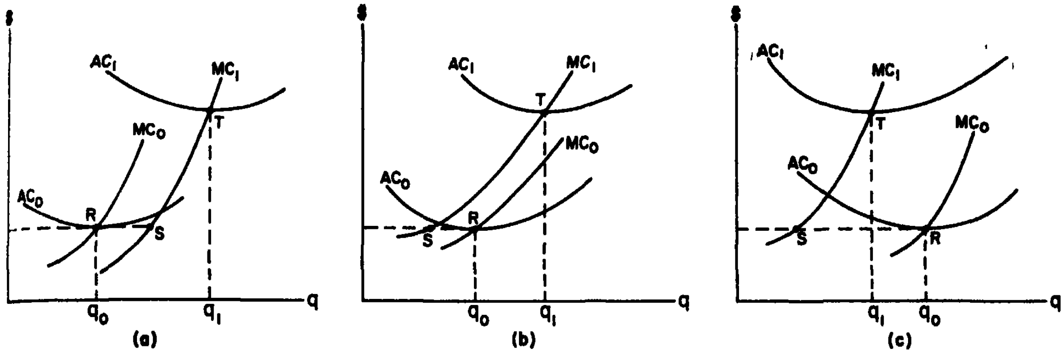


FIGURE 3

on the right-hand side of (9), as given in (8) and my Figure 2. The F-S discussion applies in full, but the underlying situation should now be somewhat clearer. Since $C''>0$, T must always be to the right of S. From (4), we know that when the ith input is inferior ($\eta_{ic}<0$), S will be to the right of R, and thus T will be to the right of R, so output must rise. But if the ith input is not inferior ($\eta_{ic}>0$), S will be to the left of R; if it is superior ($\eta_{ic}>1$), then the horizontal distance between R and S will be greater than that between S and T, so output will fall.

Finally, we briefly discuss F-S (2.7) and (2.10), which give the short-run (output changing, but output price constant) and long-run (output price varying) responses of input demand to a change in input price. Differentiating $x_j=x_j(p;\ q)$ with respect to p_i, we have

$$(11)\qquad \frac{dx_j}{dp_i}=\frac{\partial x_j}{\partial q}\frac{dq}{dp_i}+\left(\frac{\partial x_j}{\partial p_i}\right)_{dq=0}$$

Here $(\partial x_j/\partial p_i)_{dq=0}$ is just a move along an isoproduct surface (the F-S "substitution effect"). Now the question is simply what happens to q. In the short run, we have $dq/dp_i=(\partial q/\partial p_i)_{dr=0}$, and F-S (2.7) becomes

$$(12)\qquad \left(\frac{dx_j}{dp_i}\right)_{dr=0}=\frac{\partial x_j}{\partial q}\left(\frac{\partial q}{\partial p_i}\right)_{dr=0}+\left(\frac{\partial x_j}{\partial p_i}\right)_{dq=0}$$

The qualitative information which F-S obtain from (2.7) and (2.8) follows immediately from my (12), using the definition of inferiority together with our (3) and (10).

For F-S (2.10), we are in the long run, so we must substitute (7) into the right-hand side of (11):

$$(13)\qquad \frac{dx_j}{dp_i}=\frac{\partial x_j}{\partial q}\left[\frac{\partial q}{\partial r}\frac{\partial r}{\partial p_i}+\left(\frac{\partial q}{\partial p_i}\right)_{dr=0}\right]+\left(\frac{\partial x_j}{\partial p_i}\right)_{dq=0}$$

Again, it is easy to show (see the Appendix) that my (13) is precisely equivalent to F-S (2.10). Inspection of (13) together with (3) and (8)–(10) yields all the qualitative information in the F-S discussion of their (2.10), as well as a better intuitive grasp of the relationship between dx_j/dp_i and dq/dp_i in long-run competitive equilibrium.[7,8]

[7] John Black has pointed out to me that although the F-S discussion of their (2.10) (my 13) is complete and correct for the individual firm, it leaves open the question of what happens to long-run *industry* demand for an input if its price increases. As F-S say, even if long-run dq/dp_i is positive for the firm, industry output must fall in response to an input price increase, and this is achieved by an exodus of firms from the industry. Since long-run dx_i/dp_i may be positive for the firm, we should answer the corresponding question for input demand. Black has provided the following proof that in the long run, $dX_i/dp_i<0$, where X_i = industry demand for the ith input, Q = industry output, and the industry is composed of homogeneous firms: If $d/dp_i(X_i/Q)<0$, then since $(dQ/dp_i)<0$, we shall have $dX_i/dp_i<0$. Identical

firms implies $X_i/Q = x_i/q$, so $(d/dp_i)(X_i/Q) = (d/dp_i)(x_i/q) = 1/q(dx_i/dp_i) - (x_i/q^2)(dq/dp_i)$.

Substituting from (11) and (5) and rearranging, this becomes $(x_i/q^2)[(q/x_i)(\partial x_i/\partial q) - 1](dq/dp_i) + C_{ii}/q$. The first term is non-positive, from (6); and the second term is negative, since $C(p, q)$ is strictly concave in p.

[8] A paper by Lowell Bassett and Thomas E. Borcherding discussing some points covered here appeared after this note was written. They relate long-run dq/dp_i to the slope of the firm's expansion path.

Appendix

As stated in the text, the cost function technique is logically independent of and coequal to the production function; that is the import of the basic duality theorem. Thus we started with certain properties of the cost function and deduced all our results directly from these properties, making no reference whatsoever to the production function. It may nevertheless be useful to show explicitly that the equations in the text are in fact equivalent to those of F-S, although the latter are expressed in terms of expenditure elasticities, elasticities of substitution, and the determinants composed of partial derivatives of the production function. It must be stressed, however, that there is no advantage in having the results in this (seemingly more specific) form. In qualitative, comparative statics analysis of the kind involved here, we are only interested in information about signs. As we have seen, this is as readily available from the curvature properties of $C(\ )$ as it is from the signs of the Hessian determinants and their principal minors (the latter being itself equivalent to concavity of the production function). This is simply a consequence of the duality theorem.

We shall demonstrate equivalence for two of the most important pairs of equations, our (6) and F-S (21), and our (13) and F-S (2.10).

By definition, $\eta_{ic} = (\partial x_i/\partial C) \cdot (C/x_i)$. We then have

$$\eta_{ic} = \frac{\partial x_i}{\partial q} \cdot \frac{\partial q}{\partial C} \cdot \frac{C}{x_i} = \frac{\partial x_i}{\partial q} \cdot \frac{C}{C' x_i}$$

Using the fact that in long-run equilibrium, $C' = C/q$, we have

$$\text{(A.1)} \qquad \eta_{ic} = \frac{\partial x_i}{\partial q} \frac{q}{x_i}$$

From F-S (2.5), if only q changes we have $\partial\lambda/\partial q = -\lambda(F^*/F)$. Now

$$1/\lambda = C',$$

so

$$\begin{aligned} C'' &= -(1/\lambda^2)(\partial\lambda/\partial q) \\ &= (1/\lambda)(F^*/F) = C'(F^*/F) \end{aligned}$$

By hypothesis, $C' = (C/q)$, so we have

$$\text{(A.2)} \qquad C'' = \frac{C}{q} \frac{F^*}{F}$$

Substituting (A.1) and (A.2) into F-S (21) gives our (6). Substituting (A.1), (A.2), (8), and (10) into (13), then using F-S (2.1) and (2.2) and $1/\lambda = C' = C/q$, we get F-S (2.10).

REFERENCES

L. R. Bassett and T. E. Borcherding, "The Relationship between Firm Size and Factor Price," *Quart. J. Econ.*, Aug. 1970, *84*, 518–22.

C. E. Ferguson and T. R. Saving, "Long-Run Scale Adjustments of a Perfectly Competitive Firm and Industry," *Amer. Econ. Rev.*, Dec. 1969, *59*, 774–83.

W. M. Gorman, "Measuring the Quantities of Fixed Factors," in J. N. Wolfe ed., *Value, Capital and Growth*, Edinburgh 1968.

D. McFadden, "Cost, Revenue and Profit Functions: A Cursory Review," Working Paper 86, Institute of Business and Economic Research, Berkeley, Mar. 1966.

L. McKenzie, "Demand Theory without a Utility Index," *Rev. Econ. Stud.*, 1957, *24*, 185–89.

R. D. Portes, "Input Demand Functions for the Profit-Constrained Sales-Maximizer: Income Effects in the Theory of the Firm," *Economica*, Aug. 1968, *35*, 233–48.

———, "The Enterprise under Central Planning," *Rev. Econ. Stud.*, Apr. 1969, *36*, 197–212.

R. W. Shephard, *Cost and Production Functions*, Princeton 1953.

———, *Theory of Cost and Production Functions*, Princeton 1970.

H. Uzawa, "Duality Principles in the Theory of Cost and Production," *Int. Econ. Rev.*, May 1964, *5*, 216–20.

Income Taxes and Incentives to Work: Some Additional Empirical Evidence

By D. B. FIELDS AND W. T. STANBURY*

Only very infrequently in the social sciences is it possible to repeat an experiment. This paper is an attempt to repeat an important study whose results were published in this *Review* by George F. Break over a decade ago.[1] It is an attempt to measure the nature and extent of the income (incentive) and substitution (disincentive) effects of the high marginal rates of personal income tax on a group of professionals free to vary their work effort, and more knowledgeable than most individuals about their marginal rate of tax.

In the intervening years, marginal rates in the United Kingdom have declined only slightly and they remain considerably higher than those in both Canada and the United States (see Table 1). While a number of new investigations have been undertaken,[2] none serves to contradict Break's earlier results, and none are as secure methodologically as Break's study. For these reasons it was decided to repeat Break's study.

Richard Musgrave suggests that imposition of an income tax has two opposing effects:

> The disincentive or substitution effect whereby, the price of leisure having fallen, the amount of work declines.
>
> The incentive, or income effect, which moves people to do additional work to restore their previous level of disposable income. Assuming that leisure is a superior good, lower income means a lower demand for leisure. Since theory does not predict the relative strength of these two effects only empirical research can assist in establishing their nature and strength.

I. Methodology[3]

The technique for the selection of the *random* sample of solicitors and chartered accounts in and about London followed very closely the systematic sample employed by Break. The net sample of 285 consisted of 172 in the London sample and 113 in the country sample. Ninety-four of the London sample were solicitors (10 sole proprietors, 84 partners) and 78 were accountants (74 partners and 4 sole proprietors). The country sample had 70 solicitors (12 sole proprietors, 58 partners) and 43 accountants (9 sole proprietors and 34 partners).

A printed questionnaire, patterned after that of Break, was used in all interviews and the questions were always asked in the same order. The questionnaire was divided into sections, as follows:

1. personal reaction of the interviewee to his profession.
2. personal data (age, dependents, fixed financial commitments) and work history (length of time qualified, number of partners, and staff-partner ratio).
3. length of holidays (annual and other)

* The interviews for this study were carried out during 1969 by Fields of the faculty of commerce and business administration, University of British Columbia, on a year's study leave. Mr. Stanbury is a doctoral candidate in the economics department, University of California, Berkeley, and has recently joined the faculty of commerce at U.B.C. The study was made possible by the generous financial assistance of the Donner Canadian Foundation. The authors wish to express their sincere thanks to George F. Break of the University of California at Berkeley for his encouragement and advice, and to Sir Thomas Lund of the Law Society and C. Evan-Jones of the Institute of Chartered Accountants in England and Wales for their most helpful cooperation. The London Graduate School of Business Studies kindly provided library and other facilities. Finally our thanks to the hundreds of interviewees for their unfailing courtesy and cooperation.

[1] Sept. 1957, see also Break (June 1957, 1959).

[2] See studies by Rolfe and Furness, L. Buck and S. Shimmin, James Morgan et al., Robin Barlow et al., A. Chatterjee and J. Robinson, *The Graduate Appointments Register*, and C. V. Brown.

[3] A complete statement of the methodology employed is available from the authors upon request.

TABLE 1—COMPARATIVE MARGINAL TAX RATES ON SELECTED EARNED INCOMES IN THE UNITED KINGDOM, CANADA, AND THE UNITED STATES, 1968

Income After Allowable Deductions But Before Personal Exemptions			Marginal Tax Rates on Next Unit of Income Earned (to the nearest percent)					
			Single Person			Married Person with Two Children (under age 11)		
£1	$2.60 Can.	$2.40 U.S.	U.K.	Canada	U.S.	U.K.	Canada	U.S.
£ 400	$ 1,040	$ 960	20	0	14	0	0	0
800	2,080	1,920	41	15	16	20	0	0
2,000	5,200	4,800	41	26	22	41	21	15
4,000	10,400	9 ,600	53	30	28	51	26	19
6,000	15,600	14,400	64	40	36	64	40	22
10,000	26,000	24,000	79	45	50	74	45	32
20,000	52,000	48,000	91	55	60	91	55	50
50,000	130,000	120,000	91	70	70	91	70	52

Notes: Income is "assessed income" after the subtraction of allowable deductions and expenses incurred in earning income. In the case of the United Kingdom, earned income relief and the individual's personal exemption were deducted to obtain "taxable income," and the appropriate marginal tax rate was obtained from the tax schedule. In the case of Canada, the individual personal exemption ($1000) and the minimum standard deduction ($100) were deducted from assessed income to obtain taxable income. In the case of the United States, the individual's personal exemption ($600) and the minimum standard deduction were deducted from assessed income to obtain taxable income.

In this Table we have assumed that £1 in assessed income after deductions allowed by U.K. law, but before personal exemptions, is approximately equivalent to an assessed income of $2.60 (Can.) and $2.40 (U.S.) after deductions allowed by those countries, but before personal exemptions. Break made a similar assumption with the earlier exchange rates.

The 1956 U.K. marginal tax rates were slightly higher than the 1968 rates in each income range.

and hours of work (currently, previous year, and five years previously.)

4. economic status ("better or worse off") now as compared with five years ago.
5. provisions for retirement (asked only of persons 45 and over).
6. opportunities in past twelve months to accept new work (either personally or for the firm).
7. professional and total income, and estimate of marginal rate of tax.
8. questions relating to the Capital Gains Tax, the special charge and the Selective Employment Tax.
9. questions relating to the incentive or disincentive effects on work effort of high marginal rates of income tax.

By asking questions in this order, it was possible to obtain much useful information before the subject of taxation was introduced. If the interviewee raised the matter of taxation (as a number did, for instance, when asked about their economic status and retirement provisions) this was discussed, and any relevant comments were recorded. However, the questions relating to the impact of taxes were always asked in the prescribed sequence. A very careful analysis of each interview was made to eliminate questionable tax influences. Our emphasis was on actual *changes in behavior* in response to high marginal rates of tax rather than on mere verbalizations.

The overall results of the present study are compared with those of Break (Sept. 1957, Table II, p. 536) in Table 2.

Both studies indicate that a significant proportion of the individuals studied experience or state that they experience incentive or disincentive tax effects, these we have called "gross tax effects." The elimination of questionable tax influences leaves us with those who indicate that their behavior has *changed* as a result of high marginal rates of tax. The subsequent analysis in the paper will deal only with net tax effects.

While Break found that the number of

persons experiencing net tax effects was small (13 percent experienced disincentives and 10 percent incentive effects) but significantly greater than zero, he did not find the disincentive effect to be significantly greater than the incentive effects. However, we found that 18.9 percent of the sample experienced disincentive effects and only 11.2 experienced incentive effects. The difference is statistically significant at the .02 level in a single tail test. The proportion experiencing a disincentive effect has apparently increased over time. It was 13.1 percent in Break's study and 18.9 percent in the current study—the difference is statistically significant at the .03 level. Unfortunately, it is not possible to measure the *net* loss to society in terms of the quantity of labor services.[4]

II. Multiple Tax Effects

At this point in the analysis, an important distinction must be drawn. Throughout his paper (Sept. 1957), Break speaks of the number of "tax effects." Since our study indicated the presence of a substantial number of persons who experienced *multiple* net tax effects (5 of whom experienced *both* incentive and disincentive effects and are counted twice), the number of persons experiencing at least one net tax effect (86) is substantially less than the total number of net tax effects (101). On the other hand, Break records very few multiple net tax effects (5), 4 of which are both incentive and disincentive effects. Consequently he finds that 70 persons record only 71 net tax effects.[5]

[4] One attempt to do this was carried out in the context of a somewhat similar study by Robin Barlow et al.

While previous empirical evidence indicates that the absolute size of both the disincentive and the incentive effects are small (but significant) and that the two effects about offset each other, one cannot conclude that since total output is little affected there is no cause for concern with the effects of high marginal rates of taxation. As A. R. Prest (p. 265) points out: "The essential point is that we get a different product mix, through shifting of factors trying to minimize tax liabilities, from the one we should get in the absence of high marginal rates (but assuming an equal tax take in some hypothetically innocuous form)."

Table 2—Comparison of the Number of Persons Experiencing One or More Tax Effects

	Distinctive Tax Effects	Incentive Tax Effects
Fields and Stanbury	68/285 = .238	62/275 = .218
	14/285	30/285
	54/285 = .189	32/285 = .112
Break	54/306 = .176	96/306 = .314
	14/306	65/306
	40/306 = .131	31/306 = .101[a]

Note: In each case, questionable tax effects are shown and removed. Thus of our 68 individuals with disincentive effects, 14 are "questionable" leaving 54 with definite disincentive effects.

[a] In his paper, Break indicates 31 net incentive *tax effects*, but in a letter to the authors (Feb. 1970) he indicates that only 30 *persons* are involved. The adjusted ratio, 30/306 = .098, is the one strictly comparable to our results above.

The appearance of multiple tax effects in a single direction is not surprising, i.e., an individual could refuse additional work opportunities and simultaneously shift the work load within the firm so as to do less work. However the existence of both income and substitution effects simultaneously makes one doubt the consistency of such respondents. If we place these respondents in the "questionable" rather than the "net tax effect" category, the change in the results does not affect any of our conclusions.

That we found over one-fifth of the number of persons experiencing tax effects to be experiencing more than one type of tax effect is an important difference between the two studies. The fact that most of these multiple tax effects were disincentive effects reinforces our conclusion that the disincentive effect of continued high marginal rates of tax has increased over time in the United Kingdom.

[5] We found that 9 persons experienced 2 types of disincentive effects and one experienced all 3 types. Only 3 persons experienced 2 types of incentive effect. Because persons experiencing both disincentive and incentive tax effects are counted twice, the 86 persons in our study represent only 81 natural persons. For the same reason the number of natural persons in Break's study is 66.

III. Analysis of the Net Tax Effects

After eliminating all questionable tax influences we are left with 66 disincentive tax effects and 35 incentive tax effects which are shown in Table 3 by *type* of tax effect. Break's 1956 results are shown beside the current results. Not only does the current study show a greater total number of disincentive effects, but it also shows that there has been a considerable shift in the proportion of tax effects recorded in each category. While Break found that 47.5 percent of the net tax disincentive effects reported were in category 1 *Refusal of additional work opportunities*, the current study found only 18.2 percent of disincentive effects in this category. This difference is statistically significant at the .01 level. The current study finds that 45.4 percent of the disincentive consists of a *Shift of work-load within the firm* while Break found only 27.5 percent in this category, the difference is significant only at the .08 level. Break found one-quarter of the disincentive effects in category 2, *Restriction of effort to generate new work*,[6]

[6] Break's category 2 was entitled "Reduced incentives to seek new clients," (see Sept. 1957, Table III, p. 539).

TABLE 3—COMPARISON OF TYPES OF NET TAX EFFECTS[a]

Type of Tax Effect[b]		Fields & Stanbury Number	Fields & Stanbury Percent	Break Number	Break Percent
Disincentive 1	Refusal of additional work opportunities	12	18.2	19	47.5
2	Restriction of effort to generate new work	24	36.4	10	25.0
3	Shift of work load within the firm	30	45.4	11	27.5
Total Disincentive Tax Effects		66	100.0	40	100.0
Incentive 4	Postponed retirement	17	48.6	17	54.8
5	More work on a day to day basis	18	51.4	14	45.2
Total Incentive Tax Effects		35	100.0	31	100.0

[a] Because of the existence of multiple tax effects for one-fifth of those experiencing net tax effects, it is not possible to give this table accurately in terms of *persons* experiencing such tax effects by type.

[b] The following are the definitions of the tax effect categories. Tax effect categories 1, 2, and 3 are disincentive effects, and tax effect categories 4 and 5 are incentive effects.

Category 1. Refusal to accept work offered, or refusal to extend services in other ways (e.g., being more selective in the nature of work accepted, not opening other offices). Interviewee comment: ". . . have refrained from opening a branch office due to taxation; the net return is not worth the risk."

Category 2. Restriction of effort to generate new work. Interviewee comment: ". . . definitely now do not try to generate work, entirely due to tax."

Category 3. Reduction in personal work load through allocation to other staff, or reduction in hours worked. Interviewee comment: ". . . would work longer hours if there was more in it for me at the end after tax."

Category 4. Postponement of retirement, or working longer hours up to date of retirement, due in part to tax burden. Interviewee comment: ". . . would retire tomorrow if financial circumstances permitted; tax is definitely a factor in postponing retirement."

Category 5. Extending hours of work and work effort on a day-to-day basis, partly due to the burden of tax. Interviewee comment: ". . . because of tax burden, have had to take on an undue personal load of work."

TABLE 4—COMPARISON OF THE NUMBER OF ACCOUNTANTS AND SOLICITORS, LONDON AND COUNTRY, EXPERIENCING AT LEAST ONE TAX EFFECT

	Accountants			Solicitors			Total
	London	Country	Total	London	Country	Total	
Disincentive	8/78=.103	7/43=.163	15/121=.124	18/94=.191	21/70=.300	39/164=.238	54/285=.189
Incentive	8/78=.103	3/43=.070	11/121=.091	15/94=.160	6/70=.086	21/164=.128	32/285=.112

while in the current study the proportion is 36.4 percent. However this difference is not significantly different statistically. If categories 1 and 2 are grouped together (they are somewhat similar effects, but both are quite different from category 3) we find that the difference between the current results and Break's in the proportion of disincentives in these categories together is significant at the .07 level.

Only slight differences were found between the two studies in the distribution of the two types of incentives, see Table 3.

IV. Relationship of Tax Effects to Other Factors

Accountants and Solicitors

In his 1956 study, Break (Sept. 1957) found that "comparisons of the two professional groups studied showed that solicitors and accountants showed the same reactions to taxation." We find a number of differences, see Table 4.

In terms of net incentive tax effects the difference between accountants and solicitors is not statistically significant. However, the disincentive ratio was significantly greater (.01 level) for solicitors than for accountants.

This change is consistent with other evidence, elicited during the interviews, of a growing disenchantment within the profession on the part of a number of solicitors.[7] Some further evidence that might be germane is that when asked whether they considered themselves better or worse off now, as compared with five years ago, twice as many (proportionately) solicitors as accountants said they were worse off. For all accountants the disincentive ratio was greater than the incentive ratio but the difference is not statistically significant. Solicitors, however, evidenced a significantly higher (.01 level) disincentive tax effect ratio than incentive tax effect ratio.

The Geographical Factor: London and Country

For accountants and solicitors together in the Country group, the disincentive tax effect ratio is .248, but for the London group the comparable ratio is .151. This difference is statistically significant at the .025 level. For the incentive tax effect ratio we find that the London group exceeds the Country group but that the difference is significant only at the .08 level in a single-tail test. Therefore, in the case of the Country professionals, we find that the disincentive tax effect ratio far outweighs the incentive ratio—the difference is highly significant. However, for the London group, the difference is not statistically significant. While Break did not segregate tax effects between solicitors and accountants in his paper, he did report that "tax disincentives tended to occur less frequently among London respondents than among those practising in the smaller communities" (Sept. 1957, p. 544).

[7] In reply to the question "If you had a second opportunity would you still choose to become a solicitor (accountant)?", 70.6 percent of accountants replied "Yes" but only 52.8 percent of solicitors did so. When asked about the disadvantages of being a solicitor (accountant) 22.7 percent of accountants replied "none" but only 8.1 percent of solicitors gave this reply. The fact that the Prices and Incomes Board regulates conveyancing fees (an important part of solicitor's income) was frequently cited as a restrictive factor.

TABLE 5—COMPARISON OF THE SENSITIVITY OF PARTNERS AND SOLE PROPRIETORS TO TAXATION

Persons experiencing at least one net tax effect	Partners	Sole Proprietors	Total
Disincentive			
Fields & Stanbury	47/250= .188	7/35= .200	54/285= .189
Break[a]	24/217= .110	16/72= .220	40/289= .138
Incentive			
Fields & Stanbury	24/250= .096	8/35= .229	32/285= .112
Break	21/217= .097	10/72= .139	31/289= .107

[a] Sept. 1957, Table VII.

Partners and Sole Proprietors

Table 5 compares the sensitivity of partners and sole proprietors to taxation. The results are mostly consistent with those of Break, who found "that sole proprietors show a greater sensitivity to both tax incentives and disincentives than do partners" (Sept. 1957, p. 544). Using Break's data we find that sole proprietors had a significantly higher (.02 level) disincentive tax effect ratio than did partners. Making the same comparison in the current study the sole proprietors' tax effect ratio slightly exceeded the partners' ratio but the difference is not statistically significant. In the case of the incentive tax effect ratios, Break's data indicate that while the tax effect ratio for sole proprietors exceeds that of partners, the difference is not statistically significant. However, the current study shows that the sole proprietors' ratio is significantly greater (at the .01 level) than the partners' incentive tax effect ratio.

Break found that "among sole proprietors disincentives tend to exceed incentives, whereas among partners the two types of tax influences are approximately equal" (Sept. 1957, p. 544).

Our results are somewhat different. We find that for partners the disincentive tax effect ratio exceeds the incentive ratio (.188 to .096) and the difference is significant at the .01 level. Among sole proprietors, the current study finds no statistically significant difference in the disincentive tax effect ratio and the incentive ratio. The difference between Break's results and the current one for the disincentive tax effect ratio of partners is significant at the .01 level. In particular, partners are showing a far greater tendency to shift the work load within the firm (disincentive category 3, see Table 3) than they did at the time of the previous study.

The Level of Professional Income and Marginal Rate of Tax

In attempting to compare tax effects on various levels of income, and on incomes earned at different time intervals, it is necessary to take some cognizance of the real value of the monetary unit at the two points of time.[8] For this reason, in Table 6 we used different income classes in our comparison of net disincentive and incentive tax effects than those used by Professor Break in an attempt to compare equivalent levels of *real* income. Because of the range of the income classes we could only approximate comparable levels of real income.

As might be expected when marginal rates of tax are steeply progressive, the results of the study find (consistent with those of

[8] Between January 1956 and December 1968 (the period between the two studies) the United Kingdom Index of Retail Prices rose by 46%. Money incomes have risen at least as much. One attempt to measure effective tax rates on *real* income is that by A. J. Merrett and D. A. C. Monk. They use an income multiplier which "represents the *multiple* of the annual rate of inflation in the general price level which the individual had to secure as his increase in before tax income in order that his net of tax income would remain constant in real terms" (p. 95). The formula for the income multiplier is: $m = 1 - \text{average tax rate}/1 - \text{marginal tax rate}$. Using this formula for 1956 and 1968, money incomes would have to rise, on the average, by approximately 53 percent for an individual to have the same real income in 1968 as he enjoyed in 1956.

Break) that the proportion of persons experiencing disincentive effects increases as income increases. Thirty percent of the persons in our sample earning in excess of £7,000 in 1968 experienced a tax disincentive effect, whereas for those earning less than £3,000 there is a disincentive effect experienced by 9 percent. This difference is statistically significant at the .01 level. Break found that tax incentives were rather evenly distributed over the income classes. We find that this pattern has changed, with the incentive effects occurring much more frequently in the middle and lower income classes than in the higher income classes. One notable difference is observed in our £3–4,000 class which is comparable to Break's £2–3,000 class. He found that only 8 percent of persons in this class experienced an incentive effect while we find that 24 percent of persons in the comparable income class experienced incentive effects.

If we group the two lowest income classes in 1968, i.e., under £3,000 and £3–4,000, we find that the incentive tax effect ratio is more than twice that for the two highest income classes taken together. The difference is statistically significant at the .01 level. Grouping those below and above the £4,000 income level, we find that proportion above £4,000 experiencing disincentive effects is significantly greater than the proportion below £4,000 experiencing disincentive effects.

Outside Income

As one would expect, individuals with a significant amount of outside income (over £1,000, equal to U.S. $2400) are far more subject to tax disincentives than those with no outside income.

We found that 39 percent of persons with outside incomes of more than £1,000 experienced at least one net disincentive tax effect while only 15 percent of those reporting no outside income experienced a disincentive effect. The difference is significant at the .01 level.

While 11.3 percent of those with no outside income experienced at least one incentive tax effect and 7.3 percent of those with an outside income of greater than £1,000 experienced an incentive effect, the difference (which is in the direction one would expect) is not statistically significant. The data indicate that of those persons with no outside income the proportion experiencing disincentive tax effects is not significantly different (although it is slightly greater) from the proportion experiencing incentive tax effects. This is not true for those persons with outside incomes exceeding £1,000. Proportionately over five times as many such persons experience disincentive tax effects

Table 6—Relationship Between Level of Professional Income and the Proportion of Persons Experiencing One or More Net Tax Effects

		Professional Income in £ Sterling[a]			
Break[b]	1956	<2000	2–3000	3–5000	>5000
Fields & Stanbury	1968	<3000	3–4000	4–7000	>7000
Disincentive					
Break (300 cases)	1956	7/112 = .063	7/63 = .110	10/71 = .141	16/54 = .296
Fields & Stanbury (276 cases)	1968	6/65 = .092	10/58 = .172	20/99 = .202	16/54 = .296
Incentive					
Break (300 cases)	1956	11/112 = .107	5/63 = .079	7/71 = .099	7/54 = .130
Fields & Stanbury (276 cases)	1968	8/65 = .123	13/58 = .241	7/99 = .071	4/54 = .074

[a] does not include "outside" income; see text.
[b] see Sept. 1957, Table VIII, p. 545.
Note: Of our total sample of 285, 9 persons refused to state their income; of Break's total sample of 306 persons, apparently 6 persons refused to state their income.

than experience incentive tax effects, the difference is significant at the .01 level. This result is consistent with our findings that 72 percent of persons experiencing a disincentive are paying surtax, whereas only 41 percent of persons experiencing an incentive effect are paying surtax.

Fixed Commitments

One basic difference is methodology between Break's study and ours occurs in respect of the determination of fixed financial commitments. Break defined commitments as heavy or light *in terms of age and income only* (Sept. 1957, p. 545–46). On the other hand, we asked each respondent the question, "Would you say that your fixed financial commitments are light, moderate or heavy in terms of your income?", and explained that "fixed financial commitments" would include such things as school fees, mortgage payments, pension payments, and capital commitments to the firm.

Break, on the basis of his absolute (and rather arbitrary) definition found that "No respondent with relatively heavy fixed commitments in relation to his income reacted to taxation by contracting his supply of labor; all 26 of the disincentive cases occurred within the group subject only to light commitments" (Sept. 1957, p. 546, Table IX).

Based on the subjective and relative definition (and the interviewee's own estimation) we found *no* statistically significant relationship between the relative burden of fixed commitment and the incidence of incentive or disincentive tax effects.[9]

Age

The results of relating age to the number of persons experiencing tax effects indicate that disincentive effects from taxation are spread fairly evenly throughout the three age groups; under 40, 40 to 50, and over 50. On the other hand, there is a far greater preponderance of incentive effects in the over 50 age bracket than in the other two. We find that 62.4 percent of those persons experiencing at least one incentive tax effect are over age 50 while only 29.7 percent of those experiencing a disincentive effect are in the same age bracket. The difference is statistically significant at the .01 level.

This is explained by the fact that three-quarters of the persons in the over 50 bracket experiencing incentive effects were found to be in category 4, postponed retirement. Obviously, provision for retirement is a serious problem for many professional persons, and not necessarily just those in the lowest income category.[10]

Hours of Work

We found that there are no statistically significant differences between the number of hours worked by persons experiencing tax disincentives, persons experiencing tax incentives, and persons experiencing no tax effects at all. From this it would seem fair to conclude now, as Break did in 1956 that,

> Those who were cutting down their work supplies because of income taxes were still not working less, on the average, than those who were unaffected by taxation. The inference is clear then, that the tax disincentives were concentrated among respondents who were by nature more hardworking and ambitious than the average. [Sept. 1957, p. 546]

V. A Final Note

In survey research of this type one is implicitly asking a hypothetical question—"If taxes were lower (higher) would you work more or work less?" The reply to such a question can only be obtained by introspection. Since the individual hardly operates like a regression model precisely measuring the partial effects of each of a host of variables, while holding *ceteris paribus* it is the task of the researcher to try and weigh the importance of the variables under consideration. Future research seems to lie in the direction of econometrics, but the data for such work may well be derived from more comprehensive survey information. We await the application of such tools.

[9] The standard χ^2 test was used.

[10] Two-thirds of the persons experiencing incentive tax effects and over age 50 were found to have professional income in excess of £3,000 (U.S. $7,200).

REFERENCES

R. Barlow, *The Effects of Income Taxation on Work Choices*, Study No. 4 for the Royal Commission on Taxation, Ottawa, 1967.

——, H. E. Brazer, and J. W. Morgan, *Economic Behavior of the Affluent*, Washington, Brookings Institution 1966.

G. F. Break, "Income Taxes and Incentives to Work," *Amer. Econ. Rev.*, Sept. 1957, *47*, 529–49.

——, "Effects of Taxation on Incentives," *British Tax Rev.*, June 1957, 101–13.

——, "Income Tax Rates and Incentives to Work and Invest," *Tax Revision Compendium*, Vol. 3, Committee on Ways and Means, House of Representatives, Washington 1959.

C. V. Brown, "Misconceptions About Income Tax and Incentives," *Scot. J. Polit. Econ.*, Feb. 1968, *15*, 1–21.

L. Buck and S. Shimmin, "Is Taxation a Deterrent?", *Westminster Bank Rev.*, Aug. 1959, 16–19.

A. Chatterjee and J. Robinson, "Effects of Personal Income Tax on Work Effort: A Sample Survey," *Can. Tax J.*, May-June 1969, *17*, 211–20.

A. J. Merrett and D. A. C. Monk, *Inflation, Taxation and Executive Remuneration*, London 1967.

J. W. Morgan, M. H. David, W. J. Cohen, and H. E. Brazer, *Income and Wealth in the United States*, New York 1962.

R. A. Musgrave, *The Theory of Public Finance*, New York 1959, pp. 232–48.

A. R. Prest, *Public Finance*, London 1960.

S. E. Rolfe and G. Furness, "The Impact of Changes in Tax Rates and Methods of Collection on Effort," *Rev. Econ. Statist.*, Nov. 1957, *39*, 394–401.

C. T. Sandford, *Economics of Public Finance*, London 1969, p. 103 cites a survey conducted on behalf of *The Graduate Appointments Register*, April 1967.

Fiscal and Monetary Policy Reconsidered: Comment

By BENT HANSEN*

In a recent issue of this *Review*, Robert Eisner discussed the apparent inefficiency of fiscal policy on the inflationary development in the United States during the last few years. He argued, in particular, that the income tax surcharge could not have been very helpful in dampening the inflationary pressures in 1969 and that expenditure reductions would have been much more efficient. Eisner appraises fiscal policy measures exclusively according to their effects on aggregate final demand per dollar of government revenue or spending. He arrives thus at a ranking with temporary income taxes at the bottom and reductions of public purchases of goods and services at the top. This ranking is identified with a ranking according to anti-inflationary efficiency.

Eisner's reasoning leans heavily upon the life-cycle hypothesis: Since current consumption is largely determined by average expected disposable income, and since the latter cannot possibly change much as the consequence of a surcharge that is expected to last for only a single year, the surcharge could not have had much direct effect on current consumption either. Even after allowance for "normal" multiplier effects (there is little reason to believe that income-earners should perceive the temporary nature of the indirect effects), the impact on the inflationary pressures has, therefore, been small. A similar argument is applied to the effects of the surcharge on investment demand. Because expenditure cuts, even if temporary, affect demand with their full amount and with normal multiplier effects, it stands to reason that their total effects should be substantial. Eisner also mentions the well-known fact that temporary taxes on purchases of (durable) consumer and capital goods have strong direct effects on demand—precisely because they are temporary; several European countries have taken advantage of this circumstance on various occasions. Broadly based indirect taxes (general sales or value added taxes) presumably would be even more efficient for the purpose than the more narrow-based taxes mentioned by Eisner, and measured by demand effect per dollar government revenue, we may have here the most efficient short-term instrument. The possibility of financing the given increase in the war expenditures through sales of government bonds is also discussed. Although Eisner seems to be wrong (even on his own model) in maintaining that at a given money supply, the effects of bond-financed public expenditures cannot, in principle, be anti-inflationary with respect to aggregate demand, he may be right, of course, as a matter of fact.[1]

Thus, there seems to be relatively little to object to in Eisner's reasoning, as far as it goes. But it does not go very far. He is probably justified to blame the advocates of the surcharge for basing their conjectures about its aggregate demand effects upon too crude consumption and investment functions. However, he overlooks the fact that the life-cycle income hypothesis (or, for that matter, the permanent-income hypothesis) is itself a rather narrow theory of household behavior that leaves various reactions out of the picture, and that the effect on aggregate demand is not the only circumstance that

* University of California, Berkeley. My colleagues Robert A. Gordon, Abba P. Lerner, and Earl Rolph commented upon an earlier draft, and the suggestions and critical remarks of an anonymous referee caused me to make some further revisions. I am indebted to all of them. The clerical services provided by the Institute of Business and Economic Research, University of California, Berkeley, are also acknowledged.

[1] I am talking here about financing through sales of ordinary negotiable government bonds. In his concluding remarks, Eisner mentions the possibility of obtaining substantial effects from the sales of nonnegotiable savings bonds with tax or interest benefits.

matters for a ranking of fiscal (and monetary) policy measures according to anti-inflationary impact.

The life-cycle income hypothesis disregards possible short-term labor supply reactions[2] that may take the place of demand reactions and be significant for the impact on the excess demands in the system and, hence, on the tendencies for money wages and prices to increase. Perhaps more important, from a pratical point of view, are the possible effects on wage claims and profit margins, that is, the possible "cost-inflationary" effects of the surcharge and other fiscal and monetary policy measures. Eisner does not consider such effects either. A ranking of policy measures according to their effects on cost inflation may be entirely different from a ranking according to the effects on aggregate demand or excess demands in the system.

Taking into account short-term labor supply reactions, Eisner may be wrong to maintain that the surcharge has only a minor anti-inflationary impact through its effects on the demand-supply situation, although the possible labor supply reactions probably are too weak to upset his ranking of the policy measures. Cost-inflationary considerations, on the other hand, strengthen his conclusion that a reduction of other defense expenditures would have been one of the most effective ways of curbing the inflationary tendencies arising from the Vietnam War.

My first point is thus that a temporary income tax such as the surcharge will influence labor supply and may be followed by a temporary increase in the supply of labor from households affected by the surcharge. This reaction requires a relatively strong income effect and weak substitution effect on leisure. That the substitution effect should be small in the short run seems reasonable, and at least for families with institutionalized savings—and typically these are wage and salary earners—it seems likely that the short-term income effect should be strong. Considering the relatively small size of the surcharge (measured on disposable income), family members normally not in the labor force (married women, say) may not be induced, to any significant extent, to enter the labor force temporarily. But income-earners already in jobs might take extra jobs, work overtime, and so forth, and unemployed family members already looking for work might cut short their search efforts and accept jobs not considered sufficiently attractive otherwise.

To the extent that the labor supply actually increases, the anti-inflationary effects are obvious. The economy will find itself with an easier labor market and a weaker excess demand pressure on money wages. The rate of increase of money wages should thus tend to be dampened, and if price markups are triggered mainly by money wage increases (at given markup ratios), there will be corresponding dampening effects on price inflation. Moreover, if the households actually succeed in increasing their employment and, hence, their income before tax, production and supply of commodities and services will increase at an unchanged demand (assuming, in line with Eisner's reasoning, that consumption is approximately constant in any case). If the price markup ratios are flexible and influenced by excess demand for commodities, price inflation will be dampened further in this way.

Of course, Eisner may brush these arguments aside in the same way that he deals with the possible effects of general bond sales and changes in money supply—namely, by putting the burden of empirical proof on those who suggest that such effects may be strong. This, perhaps, is as it should be. Existing empirical labor supply studies do not throw much light on the concrete prob-

[2] This point has been made earlier in a more general way. In a survey of consumption theory Daniel Suits says:

> . . . during periods of prosperity and full employment, many spending units are free to offer more or less labor services and hence, within limits to determine their own incomes. The behavior of the spending unit results in the selection of both a volume of consumption expenditures *and* a level of income. . . . " Suits concludes his discussion by stating: "To the extent that steps (e.g. tax reduction) taken to raise incomes result in moderating pressures on the spending unit to produce, they may be offset by reduced earning effort. [p. 23]

lem at hand, however. Short-term changes in labor supply are certainly strong enough to have significant effect on the labor market situation, but seem to be induced by changes in unemployment rather than wages. In the present case we are confronted, however, with temporary changes in disposable income and their effects may differ from those of permanent wage changes. Be this as it may, Eisner's analysis remains incomplete as long as he does not consider labor supply effects.

My second point is that fiscal policy measures may have effects on wage claims and profit margins and thus have cost-inflationary effects that should be considered when discussing the optimal policy mix. It is convenient to start out from the empirically, relatively well-established hypothesis that the rate of change of money wage rates is partly determined by the rate of change of prices. This relationship implies that an increase in indirect consumer taxation, imposed to diminish consumer demand and thus inflation, will tend to have cost-inflationary effects that might not have arisen had the same demand curb been accomplished by other means. What is gained on the demand side may thus be lost on the cost side. Temporary, broadly based consumer taxes which, as already mentioned, may have the strongest demand effects of all fiscal policy measures, may also have the strongest cost-inflationary effects. It might be argued that a cost inflation of this type can be, at most, of limited size and duration, and if the indirect taxes are temporary, it will be reversed sooner or later. But this is beside the point. In relation to the balance of payments, for instance, even a temporary cost inflation may be dangerous; once the exchange reserves are lost, they are lost forever. *Ceteris paribus* with respect to the demand-supply situation, we shall always be interested, therefore, in minimizing possible cost-inflationary effects, whether temporary or permanent.

Thus, granted that indirect taxes imposed to curb consumer demand may have cost-inflationary effects, it is difficult to see why this should not apply to any kind of taxation that aims at curbing the consumer demand of wage-earners. Union policies may be directed toward securing a certain standard of living rather than certain money, or real wages before taxes.[3] And when wage claims arise directly from the rank and file or from the shop floor—and recent labor market events can leave no doubt that wage claims sometimes do arise in this way—they are probably generated by a strong dissatisfaction of wage-earners with respect to their *general* economic position (disregarding purely political motives). Whether the economic position of the man on the shop floor has been eroded through increasing prices, the surcharge, an increase in payroll taxes, or even an increase in loan rates may matter little to him. What does matter is the increasing difficulty his family has in making ends meet—for whatever reason. Wage claims may be triggered by anything that makes wage-earners feel worse off. If we insist upon this, we should admit, however, that anything that makes wage-earners feel better off may dampen wage claims. Public expenditures made to improve public services to wage-earners could conceivably have such an effect, particularly if they directly and obviously lead to savings for wage-earner households. Improved public health services that directly reduce the household's need for private health service might be a case in point.

The reasoning in the last paragraph applies, perhaps, mainly to spontaneous wage claims and wage actions arising directly from the shop floor.[4] We do, however, live in a time when spontaneous labor market actions seem to become more and more widespread and important. They are not just a "British disease," and policy makers may have to consider them in any country in any context. Economic theory should also try to live up to the situation.

To complete the discussion of cost-inflationary effects, we should, in principle, also take into account effects of business taxes on

[3] This point has also been made by Brennan and Auld

[4] The recent Swedish civil servant strike was a union initiated attempt to restore relative disposable salaries as a reaction to the governments' equalization policies

prices, and effects of public services on private production costs (improved road services reduce private transport costs, and so forth). There is much disagreement about the possibilities for business to shift income taxes on to prices, but at least for corporate business cost-inflation may arise in this way. The effects of public services on private costs cannot be disputed, but from a short-run point of view, this is probably a secondary matter. Finally, it might be argued that since profit margins in contracts about delivery of military equipment tend to be excessive, a cut in defense expenditures might be cost-deflationary through spill-over effect on profit margins in "civil" business.

To conclude, it would seem that the ranking of *temporary* fiscal policy measures with respect to their cost-inflationary effects differs radically from their ranking with respect to demand (or excess demand) effects. On the latter criterion, the ranking, according to effect on demand per dollar revenue or expenditure, may be: income taxes, reduction of purchase for civil or defense purposes, broadly based taxes on purchases of goods and services. On the cost-inflationary criterion, I would tentatively suggest the following ranking: reduction of defense expenditures, taxes on investment, reduction of "useful" civil expenditures, income taxes, and consumer goods taxes. The reasons for this ranking have partly been given above: reduction of defense expenditure may have a positive cost-deflationary effect, and investment taxes do not affect consumer goods prices at present (we are only interested in short-term effects). Reduction of useful civil expenditure may be felt by wage earners through the need for private expenditures. At the upper end we have, of course, income taxes and, at the top, consumer goods taxes.

Granted then, that at given demand-supply effects of alternative fiscal and monetary policy measures, inflation should be curbed by measures that have the smallest adverse effects on wage claims and profit margins, there can be little doubt that a cut in other defense expenditures, given the escalation of the Vietnam War expenditures, must rank high among the conceivable measures that could have served to improve the balance between demand and supply. A surcharge or consumption tax financed war that nobody wants to pay for could, on the other hand, through continuous wage claims and price markups, even lead to one of those permanent inflations where the participants in the social cake party together claim more than one hundred percent of the cake, and each participant administers a parameter that can restore his share each time another participant has succeeded in cutting it down (see George Akerlof). Such consequences tend to threaten in particular when the war also leads to a loosening of the social structure, thus making spontaneous labor actions both more frequent and more successful.

REFERENCES

G. A. Akerlof, "Relative Wages and the Rate of Inflation," *Quart. J. Econ.*, Aug. 1969, *83*, 353–74.

G. Brennan and D. A. L. Auld, "The Tax Cut as an Anti-Inflationary Measure," *Econ. Rec.*, Dec. 1968, *44*, 520–25.

R. Eisner, "Fiscal and Monetary Policy Reconsidered," *Amer. Econ. Rev.*, Dec. 1969, *59*, 897–905.

D. B. Suits, "The Determinants of Consumer Expenditure: A Review of Present Knowledge," in D. B. Suits et al., eds., *Impacts of Monetary Policy: Commission on Money and Credit*, Englewood Cliffs 1963.

Fiscal and Monetary Policy Reconsidered: Comment

By John H. Hotson*

I accept Robert Eisner's thesis that "... the tax surcharge should never, on basic theoretical grounds, have been considered an effective anti-inflationary device and that, given a sufficiently excessive rate of government spending, there is little that any meaningful monetary policy can do to stop inflation" (p. 898). I also share his concern that the failures of current policies will turn our fates back to "know-nothings." I am critical not so much of what Eisner says, as of what he omits.

Is the Johnson administration's desertion of the "Guideposts" in the presence of the inflationary enemy in 1966 of no value in explaining our quickened inflation since then? If the administration had escalated that particular effort, had rallied public opinion, had acquired ultimate legal sanctions against noncompliance, would not the inflation have been lessened? George Perry's findings that the guideposts had a "significant" effect on the pace of wage changes (1967, p. 903) appear to have survived all attacks to date.[1] Perry found that during 1965 and 1966 the guideposts were reducing wage increases by about 2 percent below what would otherwise be obtained. If this level of effectiveness had been maintained, wage payments would have been reduced by about $10 billion in 1968.[2] This is as large as the impact upon demand which was expected from the $10 billion surtax, an impact which Eisner argues did not materialize because the surtax did not change personal and corporate estimates of permanent income (p. 898). Moreover, wage restraint holds down the cost level; in contrast a policy which allows excessive income gains, and then tries to tax them away involves us with barn doors and stolen horses.

The "Keynesian" economists would be less discomforted by know-nothings, in my opinion, if they themselves had been closer students of Keynes. No better place for a fresh start for arriving at correct analysis can be found than in the good book *General Theory*. It is all there: the Phillips curve,[3] the guidepost prescription,[4] and Keynes' Theory of the Price Level of which so few Keynesians appear to be even aware.

> In a single industry its particular price-level depends partly on the rate of remuneration of the factors of production which enter into its marginal cost, and partly on the scale of output. *There is no reason to modify this conclusion when we pass to industry as a whole.* [p. 294] (emphasis added)

Keynes railed against the dichotomic economics of the "classicists" in which no bridge joins the theory of relative prices with

* Professor of economics, University of Waterloo.

[1] See comments of Paul Anderson, Michael Wachter, Adrian Throop and Perry's reply, 1969. Perry demonstrates in his reply that wage change behavior since 1966 are "just as the guidepost hypothesis would predict. This behavior is not predicted by any alternative hypothesis" (p. 369).

[2] Compensation of employees totaled $513.6 billion in 1968. (See Council of Economic Advisors, p. 241.)

[3] That there is a "trade off" between wages and employment is clear from the following:

> That the wage-unit may tend to rise before full employment has been reached, requires little comment or explanation. . . . there is naturally for all groups a pressure in this direction, which entrepreneurs will be more ready to meet when they are doing better business. For this reason a proportion of any increase in effective demand is likely to be absorbed in satisfying the upward tendency of the wage-unit. [p. 301]

[4] Almost the entire 1962 guidepost policy is contained in and advocated in the following:

> . . . I am now of the opinion that the maintenance of a stable general level of money wages is . . . the most advisable policy . . . There are advantages in some degree of flexibility in the wages of particular industries so as to expedite transfers . . . But the money-wage level as a whole should be maintained as stable as possible, at any rate in the short period. . . . In the long period . . . we are still left with the choice between a policy of allowing prices to fall slowly with the progress of technique and equipment whilst keeping wages stable, or of allowing wages to rise slowly whilst keeping prices stable. On the whole my preference is for the latter alternative. [pp. 270–71]

the theory of the price level. The first was made to depend on the "homely but intelligible" concepts of micro-theory, while the second depended on the quantity of money. Keynes sought to "... escape from this double life and to bring the theory of prices as a whole back to close contact with the theory of value" (p. 293). But in vain! Keynes disciples perpetuate the classical schizophrenia. Chapter 21, "The Theory of Prices," might as well not have been written, so little use is made of it by Keynesians.

To Keynes, the price level was determined by the wage level—or more inclusively, the "cost-unit" a weighted average of the rewards of the factors entering into marginal prime cost—and the scale of output (p. 302). Changes in the money supply, or changes in the level of government spending exert their influence on the price level *through* their influence on the cost-unit, and the law of diminishing returns. If the Aggregate Production Function is as near linear as some maintain, (see R. G. Bodkin, E. Kuh, T. A. Wilson and Otto Eckstein), i.e., little diminishing returns or "bottleneck" effect, this does not vitiate Keynes' analysis. Rather it would mean that the level of Aggregate Demand, or of the money supply, can have no effect on the price level except through their effects upon the cost-unit, always excepting, as Keynes did, conditions of complete full employment where "true inflation" can occur (p. 303).[5] Perhaps we should avoid this extreme view, which negates the macroeconomic importance of so much that we teach in micro-theory. However, it is probably true that inflation control is much more a matter of preventing simultaneous upward shifts in Aggregate Supply and Demand, from upward shifts in the cost-unit, than it is a mechanistic matter of diminishing returns.

Eisner's analysis seeks to demonstrate that tax and interest rate hikes are weak anti-inflationary policies because they have only minor impacts on demand. He touches on the possibility that they are perverse only in a footnote: "... higher costs of money, like sales and excise taxes and, in the long run, business income taxes as well, tend to raise prices by raising costs of production" (p. 904). Might not the price increasing effects of higher interest and tax rates outweigh their price reducing effects, given the weaknesses Eisner highlights? Should not theorists consider this fully and econometricians sift the data in search of an answer? Economists seldom consider this possibility, however.[6] Here again we see the schizophrenia that afflicts economic theory. In the theory of the firm we discuss interest and taxes as not unimportant costs of production,

[5] We approached, but at no time achieved such "full" employment in the Vietnam expansion. At all times there was a margin of unemployed resources and government and private demands expanded together. This weakens Eisner's argument regarding the impotence of monetary policy (p. 902). A few figures: in 1968 some 2,817,000 were unemployed out of a civilian labor force of 78,737,000, or 3.6 percent. Six hundred thousand more could have been employed before the unemployment rate would have equalled the 2.9 percent recorded in 1953, and 1,872,000 more would have had to be employed to equal the 1.2 percent rate of 1944 (see Council of Economic Advisors p. 252).

[6] Exceptions would include S. Weintraub (p. 149), P. Davidson, G. Brennan and D. A. L. Auld, whose analysis ".... demonstrates the possibility that the sales tax may be subject to cumulative shifting: first shifted forward to customers ... then ... back on firms in higher income claims to match the higher cost of living—then forward ... and so on" (see Brennan and Auld, p. 525). A formal argument to support the conventional assumptions is seldom presented. However, see Earl Rolph regarding the price *decreasing* effect of higher sales taxes and G. Horwich on why interest hikes are non-inflationary. (See also my comment and Horwich's reply.) Richard Musgrave held that Rolph's conclusions are only justified in the classical model of quantity theory and pure competition, and that in the real world of imperfect competition, forward shifting will be the "typical case" (1953, p. 514). In his well-known text, Musgrave treats of the classical case (1959, pp. 364–71) but concludes that in the real world, "... A potential inflationary gap ... may be closed by an increase in consumption taxes only while permitting some rise in the price level, reflecting the increase in cost due to tax" (1959, pp. 447–48). He holds, however, that this tax push inflation is "... a once-and-for-all increase, to be distinguished from the continuous increase that results if the inflationary gap is not closed," (p. 447) thus missing the possibilities highlighted by Brennan and Auld. Musgrave's conclusion that the Corporate profit tax is also fully shifted, (see Marion Krzyzaniak and Musgrave 1963) likewise has tax push implications, as his critic Robert J. Gordon points out (1967, p. 732). However, neither carry the point further in their subsequent exchange.

and in the theory of tax incidence we depict the taxed businessman attempting to shift his taxes forward or backward, but we fail to develop the macroeconomic implications of all this.

The theory that raising the tax and interest payments of business, and thereby causing some unemployment, reduces inflation involves the assumption of backward shifting to taxes and interest payments to factor suppliers. This is certainly possible and probably occurs to some extent over time, but that this is the sole, or even the major, impact of these policies upon incomes and prices appears highly dubious. Given the organization of producer interest groups and the failure of consumers to organize, forward shifting via higher prices appears the more likely outcome.

The theory that tax and interest rate hikes are inflationary in net effect is somewhat analogous to the balanced budget multiplier theory, but harder to quantify.[7] Increased government spending is very expansionary if it is financed by deficits and new money creation. It is still expansionary, but less so, if it is financed by higher taxes. Similarly, increased government expenditures, (in a nearly fully employed economy) are highly inflationary if they are financed by deficits and new money creation. Are they not still inflationary, but less so, if they are financed by higher taxes and the rate of interest is allowed to rise?

Rather than turn the world over to know-nothings, Keynesians should cross over the bridge Keynes built for them.

[7] Much would seem to depend upon which tax is increased. Similarly, borrowers differ greatly in their ability to pass on their interest costs by increasing the price of whatever it is they sell. In a world of strong unions, however, it would seem that even the personal income tax is shiftable with a lag. For some discussion of these matters and attempts at quantification, see Hotson (1967).

REFERENCES

P. S. Anderson, "Wages and The Guideposts: Comment," *Amer. Econ. Rev.*, June 1969, *59*, 351–54.

R. G. Bodkin, "Real Wages and Cyclical Variations in Employment: A Re-Examination of the Evidence," *Can. J. Econ.*, Aug. 1969, *2*, 353–74.

G. Brennan and D. A. L. Auld, "The Tax Cut as an Anti-Inflationary Measure," *Econ. Rec.*, Dec. 1968, *44*, 520–25.

P. Davidson, "Rolph on the Aggregate Effects of a General Excise Tax," *Southern Econ. J.*, July 1960, *27*, 37–42.

R. Eisner, "Fiscal and Monetary Policy Reconsidered," *Amer. Econ. Rev.*, Dec. 1969, *59*, 897–905.

R. J. Gordon, "The Incidence of the Corporation Income Tax in U.S. Manufacturing, 1925–62," *Amer. Econ. Rev.*, Sept. 1967, *57*, 731–58.

———, "Incidence of the Corporation Income Tax: Reply," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1360–67.

G. Horwich, "Tight Money, Monetary Restraint, and the Price Level," *J. Finance*, Mar. 1966, *21*, 15–33.

———, "Tight Money as a Cause of Inflation: Reply," *J. Finance*, Mar. 1971, *24*, 156–57.

J. H. Hotson, "Neo-Orthodox Keynesianism and the 45° Heresy," *Nebr. J. Econ. Bus.*, autumn 1967, *6*, 34–49.

———, "Tight Money as a Cause of Inflation: Comment," *J. Finance*, Mar. 1971, *24*, 152–56.

J. M. Keynes, *The General Theory of Employment, Interest and Money*, London 1936.

M. Krzyzaniak and R. A. Musgrave, *The Shifting of The Corporation Income Tax*, Baltimore 1963.

——— and ———, "Incidence of the Corporation Income Tax: Comment," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1358–60.

E. Kuh, "Unemployment, Production Functions, and Effective Demand," *J. Polit. Econ.*, 1966, *74*, 238–46.

R. A. Musgrave, "General Equilibrium Aspects of Incidence Theory," *Amer. Econ. Rev. Proc.*, May 1953, *43*, 504–17.

———, *The Theory of Public Finance*, New York 1959.

G. Perry, "Wages and the Guideposts," *Amer. Econ. Rev.*, Sept. 1967, *57*, 897–904.

———, "Wages and the Guideposts: Reply,"

Amer. Econ. Rev., June 1969, *59*, 365–70.
E. A. Rolph, *Theory of Fiscal Economics*, Berkeley 1954.
A. W. Throop, "Wages and the Guideposts: Comment," *Amer. Econ. Rev.*, June 1969, *59*, 358–65.
M. L. Wachter, "Wages and the Guideposts: Comment," *Amer. Econ. Rev.*, June 1969, *59*, 354–58.
S. Weintraub, *An Approach to the Theory of Income Distribution*, Philadelphia 1958.
T. A. Wilson and O. Eckstein, "Short-Run Productivity Behavior in U.S. Manufacturing," *Rev. Econ. Statist.*, Feb. 1964, *48*, 41–54.
Council of Economic Advisors, *Economic Report of the President*, Washington, Jan. 1969.

Fiscal and Monetary Policy Reconsidered: Comment

By BARBARA HENNEBERRY AND JAMES G. WITTE*

According to Robert Eisner in a recent issue of this *Review*, the keeping of a tight rein on the money supply can be expected to be an inadequate defense against inflation in the presence of a rise in government spending.[1] He then postulates a situation wherein government spending rises and taxes are not increased sufficiently to counter the expansionary effects on aggregate demand of this increment in government purchases. After further postulating that the nominal quantity of money is held constant, Eisner then points out that if the demand for real money is negatively related to the rate of interest, the velocity of circulation will increase. Under the assumed conditions of full employment, this will result in a rise in the price level. Thus, Eisner concludes, contemporary economic theory implies limitations on monetary policy (as well as tax policy) in combatting inflation.

Two points are worthy of note, and they both constitute important implications of the "modern" economic theory so dear to Eisner's heart. The first point is that as long as real commodity demand is a decreasing function of the rate of interest, there is some reduction in the quantity of money which would close the inflationary gap. Monetary policy is not logically constrained to zero or positive changes in the quantity of money. To use the terminology of J. R. Hicks and Alvin Hansen, the higher the (absolute) value of interest-rate elasticity of the *IS* function, the smaller need be such reduction, but there is some reduction which will offset rise in velocity induced by the increase in aggregate demand. Monetary policy, then, need not be rendered ineffective by the sheer existence of an interest-elastic *LM* function as Eisner seems to imply. After all, velocity is a function of the interest rate in any Keynesian model, but that does not imply that some reduction in the money supply could not serve to offset an interest-induced rise in velocity.

Second, a given shift in aggregate demand, with a constant supply of money, will produce a *one-time* increase in velocity and prices. Yet, as Eisner correctly points out, the problem during the 1967–69 period was that of a continuing rise in prices which has lasted well beyond the gradual elimination of the fiscal stimulant. Thus, in order to explain the continuing rise in prices one must account for a further exogenous disturbance of a more continuing nature. We submit that the continuing disturbance could be the employment of a monetary policy which permitted the money supply to rise steadily at quarterly rates which tend to exceed the growth in real *GNP*.[2] While the effect of the fiscal stimulant is to produce a new, higher equilibrium value of the rate of interest which closes the inflationary gap, the attainment of such an equilibrium rate of interest

* Former assistant professor of economics, Ball State University and professor of economics, Indiana University, respectively.

[1] The authors have chosen to comment only upon Eisner's monetary statement; they are in substantial agreement with him on his fiscal policy position.

[2] The referee has indicated to us his belief that our test of the direction of monetary policy, viz. the rate of growth of the money supply relative to the growth of real *GNP* may be inappropriate. He suggests, instead, that the relevant test of the tightness of a monetary policy in an inflationary situation "is not whether the money supply increases more rapidly than the growth in real *GNP* but whether it increases more rapidly than the growth in the money value of *GNP*"; by this criterion, he says, "monetary policy was in the proper direction but unsuccessful." We must disagree with this definition. The ratio of nominal *GNP* to the money supply is equal to the velocity of circulation. Use of such a criterion would mean that a tight money policy is in effect *whenever* velocity rises. We would then have to characterize monetary policy in the 1946–51 period as tight, despite the fact that the Treasury Bill rate was pegged by continuous open-market operations. By our definition, monetary policy during the early postwar period was clearly one of excessive ease; it is our contention that, in 1946–51 as in 1966–68, excessively easy monetary policy led to substantial increases in the general price level.

can be prevented by expansionary behavior on the part of the central bank. Central bank action which inhibits the tendency of the rate of interest to rise to its inflationary-gap-closing levels results in a positive rate of inflation that continues at least as long as the growth in the money supply continues, and this, we submit, is what has constituted the phenomenon exhibited over the time period in question. Contrary to Eisner, the inflationary problem results, not from the *inadequacy* of monetary policy but from the *improper use* of monetary policy.

REFERENCE

R. Eisner, "Fiscal and Monetary Policy Reconsidered," *Amer. Econ. Rev.*, Dec. 1969, *59*, 897–905.

Fiscal and Monetary Policy Reconsidered: Comment

By KEITH M. CARLSON*

Robert Eisner has recently entered the debate on the relative potency of monetary and fiscal actions. He demonstrates the ineffectiveness of the 1968 tax surcharge in checking inflation, then goes on to assert that tight money would be similarly ineffective. This paper considers Eisner's analysis as it pertains to the inadequacy of monetary policy. First, it is shown that his conclusions do not necessarily follow from his own model. Second, using parameter estimates representative of other studies, it is demonstrated that Eisner's conclusions are not substantiated by the empirical evidence.

I. Solving Eisner's Model

Eisner summarizes his model of aggregate demand as follows:

Private Commodity Demand

(1) $$C = C(Y, Y^e, i, A, M)$$

Real Money Demand

(2) $$M^D = M^D(X, i, A)$$

Output Equilibrium

(3) $$X = C + G$$

Money Market Equilibrium

(4) $$M = M^D$$

Money-Price Identity

(5) $$M = \frac{Q}{P}$$

The symbols are defined as:

C = real consumption plus real investment demand
Y = current real income after taxes
Y^e = expected future real income after taxes
i = real rate of interest[1]
A = net real value of nonmonetary assets held by private sector
M = real cash balances held by private sector
M^D = real cash balances demanded by private sector
X = real output
G = real government demand for goods and services
Q = nominal quantity of money
P = commodity price index

This is a system of five equations with C, i, M, M^D, and P endogenous; Y, Y^e, X, G, Q are considered exogenous. To formally close the model (which Eisner does not do) another equation is added:

Nonmonetary Asset Identity

(6) $$A = \frac{1}{iP}a$$

In this equation, a is exogenous and represents the number of nonmonetary assets, each of which can be considered as a perpetuity paying one dollar per year. Eisner places considerable emphasis on the effects on A of changes in i and P without explicitly specifying a relation like (6).

Differentiating equations (1)–(6) totally yields the following:

(7) $$dC = C_Y dY + C_{Y^e} dY^e + C_i di + C_A dA + C_M dM$$

(8) $$dM^D = M_X dX + M_i di + M_A dA$$

(9) $$dX = dC + dG$$

(10) $$dM = dM^D$$

(11) $$dM = \frac{1}{P}dQ - \frac{Q}{P^2}dP$$

* Federal Reserve Bank of St. Louis. The views presented here do not purport to represent those of the Board of Governors of the Federal Reserve System or the Federal Reserve Bank of St. Louis.

[1] Eisner does not identify his interest rate as nominal or real. Since price expectations are not mentioned in the article, he apparently assumes that the nominal and the real rate are identical.

$$(12)\qquad dA = \frac{1}{iP}da - \frac{a}{i^2P}di - \frac{a}{iP^2}dP$$

To solve the system, dY, dY^e, and dX are all set equal to zero in accordance with the full employment assumption. Two cases are considered with respect to the method of financing an increase in government expenditure. The first is bond financing, or $dG=1/iP\ da$ (thus $dQ=0$); the second is money financing, or $dG=1/P\ dQ$ (thus $da=0$).[2]

Case 1: Bond Financing

To show the effects of bond financing on the price level, the system is reduced to the following two equations:

Commodity Equilibrium

$$(13)\qquad \left(\overset{(+)}{C_A \frac{a}{i^2P}} - \overset{(-)}{C_i}\right)di = -\left(\overset{(+)}{C_A \frac{a}{iP^2}} + \overset{(+)}{C_M \frac{Q}{P^2}}\right)dP + \overset{(+)}{(1 + C_A)}dG$$

Money Equilibrium

$$(14)\qquad \left(\overset{(+)}{M_A \frac{a}{i^2P}} - \overset{(-)}{M_i}\right)di = \left(\overset{(+)}{\frac{Q}{P^2}} - \overset{(+)}{M_A \frac{a}{iP^2}}\right)dP + \overset{(+)}{M_A}dG$$

The signs of the partial derivatives (or terms including levels of variables), as postulated by Eisner, are written above those derivatives. Clearly, the commodity equilibrium equation, when charted in the (i, P) plane (following Don Patinkin), is negatively sloped. The money equilibrium curve may be either positively or negatively sloped. Conventionally, it is assumed to be positively sloped, but if $M_A(a/iP^2) > Q/P^2$ (i.e., a strong real nonmonetary asset effect on real money demand), the money equilibrium curve is negatively sloped.

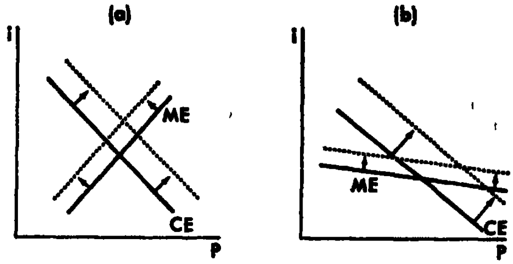


FIGURE 1. CASE 1: BOND FINANCING OF GOVERNMENT SPENDING

The effect of a change in government spending financed by bond issue is shown in the two panels of Figure 1. Panel (a) shows the effect of an increase in government spending for the positively sloped ME subcase. For an increase in government spending financed by bond issue, the effect on i is unambiguously positive, while the effect on P is ambiguous, depending on the relative magnitude of shifts of the two curves. Panel (b), with a negatively sloped ME curve, shows that the effect of an increase in government spending is ambiguous for both i and P.

As a result of solving the model, there is no basis, on strictly logical grounds, for accepting Eisner's conclusion that an increase in government spending, with the nominal money stock held constant, leads to an increase in the price level.[3] Any such conclusion requires consideration of the empirical evidence.

Case 2: Money Financing

Since Eisner goes on to argue that monetary policy would be ineffective in combatting inflation, a second case is considered before examining the evidence. Rather than examine the effect of a decline in the money stock, the effect of an increase in the money

[2] These two cases indicate that only two of the three policy variables, dG, da, and dQ, can be determined independently; thus the third becomes endogenous in the model. The importance of this government budget restraint has been emphasized by Carl Christ.

[3] Strangely enough, Eisner shows that he is aware of these possibilities. But the discussion of these possibilities is relegated to a footnote.

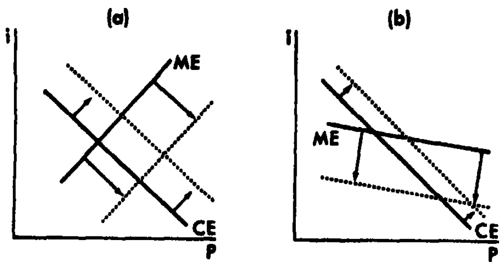


FIGURE 2. CASE 2: MONEY FINANCING OF GOVERNMENT SPENDING

stock on the price level is analyzed instead. Case 2 involves an increase in government spending accompanied by monetary expansion.

The system (7)–(12) is solved as before, except that $dG=1/P\,dQ$ and $da=0$. This case reduces to the following:

Commodity Equilibrium

$$(15)\quad \left(\overset{(+)}{C_A \frac{a}{i^2P}} - \overset{(-)}{C_i}\right) di = -\left(\overset{(+)}{C_A \frac{a}{iP^2}} + \overset{(+)}{C_M \frac{Q}{P^2}}\right) dP + \overset{(+)}{(1 + C_M)}\,dG$$

Money Equilibrium

$$(16)\quad \left(\overset{(+)}{M_A \frac{a}{i^2P}} - \overset{(-)}{M_i}\right) di = \left(\overset{(+)}{\frac{Q}{P^2}} - \overset{(+)}{M_A \frac{a}{iP^2}}\right) dP - dG$$

Equations (15) and (16) are the same as (13) and (14) except for the coefficients on dG. The effects of a change in government spending financed by money issue are shown in Figure 2.

The subcase shown in panel (a) demonstrates that an increase in government spending accompanied by monetary expansion is unambiguously positive in its effect on P. The effect on the interest rate now becomes ambiguous. Likewise for the negatively sloped ME subcase, government spending financed by money issue yields an unambiguously positive effect on P. And, for this subcase, the effect on interest rates is unambiguously negative.

After examining four subcases reflecting different sets of assumptions, the conclusion follows that the effect of government spending on prices depends critically on whether or not it is accompanied by monetary expansion. If the money stock is held constant and government spending is increased, there *may* be an increase in prices, but it is not until the assumption of unchanged money is relaxed that the effect becomes unambiguous.

II. Some Empirical Evidence

Eisner's conclusions about the factors underlying inflation do not hold even within the framework of his own model. Nevertheless, it is of interest to examine the magnitudes involved. To do so, parameter estimates from a recent econometric study are used (see the paper by Carlson and Denis Karnosky).[4]

The values assigned to the parameters and relevant variables are as follows:

Impact	Steady State	Levels
$C_i = -2.55$	$C_i = -7.07$	$a/iP = 1020.00$
$C_A = .13$	$C_A = .40$	$i = 3.87$
$C_M = .87$	$C_M = 2.67$	$P = 1.18$
$M_i = -1.96$	$M_i = -20.40$	$Q/P = 151.00$
$M_A = .02$	$M_A = .21$	

The levels of variables are for late 1967. The parameter estimates are based on regressions for the 1953–67 period. Both impact and steady state coefficients are shown.

Inserting these values into the relevant equations for the two cases, the following results are obtained:[5]

[4] The study from which these estimates were taken was based on a Keynesian income-expenditure model. Parameter estimates conform with those of other studies (see Christ).

[5] The magnitudes of the coefficients were such that the stability conditions were satisfied in all cases.

	Impact Effect	Steady State Effect
Case 1:		
$dP/dG\|_{dQ=0}$	.0013	.0018
$di/dG\|_{dQ=0}$	.0228	.0014
Case 2:		
$dP/dG\|_{da=0}$	.0088	.0085
$di/dG\|_{da=0}$	−.0027	−.0193

Because the magnitudes of the multipliers are small, the difference in effect on prices is not immediately evident. However, the bond financing case shows a .11 percent impact effect on prices, while the money financing case represents a .75 percent increase (where both percents are calculated from a late 1967 base). The price effects are of the same relative magnitude for the two cases in the steady state.

These results can be applied in a similar way so as to shed light on Eisner's conclusion. From fourth quarter 1965 to fourth quarter 1969, real government purchases of goods and services increased $15.5 billion. Treating the increase for that period as a once and for all increase of half that amount in late 1965, and applying the steady-state multiplier for the bond financing case to that figure of $7.8 billion, yields an increase in the price level of .014, or 1.26 percent, from late 1965.[6] The actual increase was .19, or 17.04 percent. Though apparently correct in direction, Eisner's strong conclusion about the inflationary impact of government spending (when financed by bond issue) receives little support from the evidence.

[6] This procedure is approximate, but the contribution of deficit-financed increases in government expenditures to the explanation of changes in the price level seems small enough (less than 10 percent) to justify rejection of Eisner's main conclusion.

III. Conclusion

Eisner concluded that tax policy and monetary policy were ineffective as tools for fighting inflation. His conclusions on tax policy may be valid; this paper focuses on monetary policy, with special emphasis on its association with financing of government expenditures. Using Eisner's model, it was found that his conclusions do not necessarily follow. Since any definite conclusions ultimately depend on the values of the partial derivatives, a set of parameter estimates was used to calculate the effects on the price level of an increase in government spending financed by bond issue and by money issue. These calculations, though admittedly approximate, indicated that government spending has a much smaller effect on the price level if unaccompanied by monetary expansion. Eisner appears to be correct in the direction of the effect of deficit-financed expenditures on the price level, but empirical evidence indicates the magnitude of the effect is very small.

REFERENCES

K. Carlson and D. Karnosky, "The Influence of Fiscal and Monetary Actions On Aggregate Demand," working paper no. 4, Federal Reserve Bank of St. Louis, Mar. 1969.

C. Christ, "A Short-Run Aggregate Demand Model of the Interdependence and Effects of Monetary and Fiscal Policies with Keynesian and Classical Interest Elasticities," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 434–43.

R. Eisner, "Fiscal and Monetary Policy Reconsidered," *Amer. Econ. Rev.*, Dec. 1969, *59*, 897–905.

D. Patinkin, *Money, Interest, and Prices*, 2d ed., New York 1965.

Fiscal and Monetary Policy Reconsidered: Reply

By ROBERT EISNER*

The four comments fall into two groups. Bent Hansen and John Hotson offer little or no objection to my critique of monetary policy. As to recent and conventional fiscal policy, they also apparently share my reservations but have some strictures of their own to add. Keith Carlson, James Witte and Barbara Henneberry seem content with the rebuke to fiscalists but seek to raise very considerable objection to corresponding questioning of monetary policy. One may infer that, for perhaps varied reasons, all contributors are pleased to have fiscal policy raked over the coals, but affront to monetary gods is offensive to some of them.

I have little quarrel with Hansen's arguments and have indeed developed several of his points elsewhere (1971). I do, as he anticipates, rather doubt a substantial negative elasticity of the supply of labor with respect to real after-tax income, but will await with interest the results of any new empirical studies of this subject. I believe there is considerable substance to his suggestion that supply functions for both labor services and private products are such that higher taxes may have serious cost-inflationary effects and that certain kinds of government expenditures may be cost-deflationary. And I am similarly appreciative of Hotson's arguments that tax and interest rate increases are cost-inflationary, but more skeptical of the wisdom and efficacy of "guidepost" restraint on wages.

With my monetarist critics I have more argument. First, Witte and Henneberry are of course correct in arguing that a demonstration that holding the quantity of money constant may not be sufficient to prevent price inflation "does not imply that some reduction of the money supply could not serve to offset an interest-induced rise in velocity." It does not logically follow, however, that "as long as real commodity demand is a decreasing function of the rate of interest, there is some reduction in the quantity of money which would close the inflationary gap." This, after all, depends upon how rapidly a decreasing function we have and just how large an inflationary gap must be closed. One may even conceive a situation in which the government insists upon keeping real commodity demand at an inflationary level while reducing the quantity of "money," however it was originally defined, to zero. The consequence might then be continued inflation, with non-banking institutions furnishing debt instruments which could serve as substitutes for the money which the Witte and Henneberry-dominated Fed might have taken out of circulation.

I can only agree with the anonymous referee who points out that the test of tightness of monetary policy must involve the relation of the quantity of money to the money value rather than the real value of *GNP*. Witte and Henneberry, along with a surprising number of monetarists, strangely seem to ignore that the nominal quantity of money is most essentially an endogenous element of our monetary system as well as most if not all others. Money is allowed to vary largely with "the needs of trade," and will certainly have a very strong tendency to rise whenever the money value of *GNP* rises. Unless the quantity of money increases more than the money value of *GNP*, Witte and Henneberry (and others) would clearly have their cause and effect reversed. This partial bending of the monetary authority to the needs of an inflationary economy, the product of repeated lessons of the consequences of short-run liquidity crises, may be considered as one of the elements in a lag process that could have been expected to contribute to "a continuing rise in prices which has lasted well beyond the gradual elimination of the fiscal stimulant."

* Northwestern University and National Bureau of Economic Research. The author is grateful for comments by Robert W. Clower on a draft of this reply.

We may note that Witte and Henneberry state that "they are in substantial agreement with [Eisner] on his fiscal policy position." Carlson's similar, if not quite definitive, acquiescence in my fiscal critique—[Eisner's] "conclusions on tax policy may be valid"—puts him in a position he may well wish to reconsider. For Carlson espouses an argument, which has recently gained considerable currency if far from universal acceptance, that there is one particular kind of paper, variously defined as legal tender, properly signed and valid checks and, in some instances, properly valid pages of savings account books, which is all-decisive. Other pieces of paper, no matter how much they look like the magic ones which can do all, have little or nothing to do with inflation. Carlson's analysis of the relative effects of bond and money-financed deficit spending along with his concession to my critique of fiscal policy bring some of these views into a clearer, perhaps glaring light.

For what Carlson is saying comes down to the following. If the government increases expenditures and raises money for these expenditures by taxes for which the private sector gets nothing in return other than pieces of paper called tax receipts, aggregate demand (and at least under conditions of full employment, prices) will rise. If the same increase in government expenditures is accompanied not by increases in pieces of paper marked tax receipts but by pieces of paper called money, Carlson would look for a much more inflationary impact—and here most of us would concur. But if the increase in government expenditures involves leaving the public not with more tax receipts and not with more money but with more government bonds, Carlson develops grave doubts as to their inflationary impact. One can indeed concede that having a thousand dollar bill from the government will be more conducive to private demand than having a thousand dollar tax receipt. But is a thousand dollar bond so much worse than a thousand dollar bill as to be worse than a thousand dollar tax receipt? Carlson concedes that increased government expenditures, which leave the private sector with no more net assets denominated in monetary terms, can be inflationary. How then can he deny that similar expenditures, which leave the private sector with monetary assets that are deferred obligations of the government rather than an instant obligation, will somehow be deflationary? Indeed, since neither Carlson nor I specify the duration of the government debt issues used to finance a budget deficit, Carlson is left in the strange position of arguing that a thousand dollar noninterest-bearing bill, which he would call money, is very inflationary but a thousand dollar interest-bearing bill which may differ from the former only in that it will not become money until tomorrow, and offers a premium of sixteen cents for the delay, may be deflationary!

Carlson insists (and Hansen also notes) that my model does not absolutely preclude a decline in prices with bond-financed deficit spending. What I did was to indicate the extreme behavioral assumption necessary for this result. (" . . . it would be to argue that the increase in interest rates necessary to reduce commodity demand despite the increase in the net value of assets would not be sufficient to reduce the real demand for money, implying a shift from commodities to money as interest rates and bond holdings rise" p. 903.) Carlson's further formalization of my model along with his own rough empirical estimates may indeed be used to substantiate the argument I advanced. Carlson points out, setting up commodity equilibrium and money equilibrium curves, that effects on price are "ambiguous, depending on the relative magnitude of shifts of the two curves." What this comes down to in the case of bond financing is the question of whether, setting $dP=0$, di/dG in Carlson's equation (14) is greater than di/dG in equation (13). This is to say that

$$\frac{dP}{dG} \gtrless 0 \quad \text{as}$$

$$\frac{1 + C_A}{C_A \dfrac{a}{i^2P} - C_i} \gtrless \frac{M_A}{M_A \dfrac{a}{i^2P} - M_i}$$

Thus, as indicated qualitatively in my original paper, for deflationary consequences of debt-financed government spending, M_A must be so large and the absolute value of M_i so small as to bring about "a shift from commodities to money as interest rates and bond holdings rise."

Carlson's own empirical estimates suggest how far from this strange condition we are. For substituting the parametric estimates he furnishes we have for the critical inequality:

$$\frac{1.13}{.13\left(\frac{1020}{3.87}\right)+2.55} \overset{?}{>} \frac{.02}{.02\left(\frac{1020}{3.87}\right)+1.96}$$

or .3007 > .0028 for the "impact" result, and

$$\frac{1.40}{.40\left(\frac{1020}{3.87}\right)+7.07} \overset{?}{>} \frac{.21}{.21\left(\frac{1020}{3.87}\right)+20.40}$$

or .0124 > .0028 for the steady state, indicating that in either case bond-financed increases in government expenditures are inflationary.

Indeed Carlson's estimates suggest that if the demand for money were perfectly inelastic, that is, $M_i=0$, the effect of increased government debt in the hands of the public would still not raise the demand for money sufficiently to lower prices. Aside from any other measurement or statistical problems in the unpublished manuscript which Carlson cites for his estimates, I might point out that his a/iP is in fact the money value of total private wealth. The elasticity of demand for money with respect to total wealth is likely to yield a higher figure for M_A than the correctly estimated derivative of the demand for money with respect to holdings of government bonds, a closer substitute for money. One may conjecture that I conceded too much in even allowing $M_A>0$; M_A, at least with respect to short-term bills, may well be negative. One can hardly entertain seriously the implication of Carlson's position, contrary to the empirical results he cites, that M_A is both positive and so large as to make bond-financed deficit-expenditures deflationary when similar tax-financed expenditures are inflationary.

Another way of exposing the remarkable role ascribed by Carlson (and others) to money as opposed to other financial assets is to compare the critical last terms of his money equilibrium equations (14) and (16). For in (16), the money-financing case, there must be an adjustment of interest rates and prices to compensate for $-dG$, the reduction in the excess demand for money brought on by its increased supply. In (14), the bond-financing case, there must be an adjustment to a presumed increase in excess demand for money equal to $M_A dG$. One may ask indeed how similar bonds must become to money before M_A in (14) approaches the value of -1 in (16), and C_A approaches C_M, in the last terms of (13) and (15). Suppose those one thousand dollar bills alluded to earlier were debt instruments maturing five minutes after they were received? Sophisticated monetarists may try to deny this, but their critical implicit notion in making money sovereign is that it is somehow unique, *both* as a means of immediate exchange and in all the other qualities attributed to financial assets. Hence, adding other financial assets can never free money for more of its presumably unique role in transactions, and there can be no substitution of other financial assets which would permit the same amount of money to finance an increased dollar volume of transactions.

It is difficult to respond to Carlson's admittedly "approximate" procedure for estimating the amount of inflation that might have occurred had the quantity of money been kept constant from 1965 to 1969, since the underlying model and basis for his estimates are not presented in his paper. It may be pointed out, however, that the figures furnished by Carlson suggest that even if increased government expenditures had been financed entirely by the creation of money, presumably his case 2, the actual increase in prices from 1965 to 1969 would have been only .0092/.0019 times the 1.35 percent increase which he argues would have taken place if the quantity of money had been kept

constant (his case 1). Thus Carlson's numbers suggest that even the most inflationary financing of increased government expenditures would have accounted for an inflation of only about 7 percent, as against the 17 percent acknowledged by Carlson as the increase over this period. Since financing larger government expenditures entirely by an increase in the quantity of money would actually have been more inflationary, by both Carlson's reasoning and mine, than what actually occurred, there is clearly something lacking or wrong in Carlson's model or numbers.

I return to my original thesis. The countercyclical fiscal policy advanced during the 1965–69 period, let alone that actually adopted, should not have been expected to be successful against the war-induced inflation. And just as surely no reasonable monetary policy, even including the perhaps unreasonable policy of trying to keep the nominal quantity of money constant, would have been successful either.

REFERENCE

R. Eisner, "What Went Wrong?", *J. Polit. Econ.*, May/June, 1971, *79*, 629–41.

More on an Empirical Definition of Money: Note

By David T. Hulett*

This study was conducted to evaluate George Kaufman's extension of the Friedman and Meiselman technique for an empirical definition of money. This method defines as money that financial aggregate which satisfies two criteria: 1) it exhibits the highest correlation with *GNP*, and 2) the correlations between *GNP* and each of the components considered separately do not exceed that between *GNP* and the aggregate. The components are thus substitutes—the public alters the composition of its portfolio due to changes in supply conditions, while keeping its portfolio size constant relative to *GNP*. (See Friedman and Schwartz, ch. 2; and J. R. Hicks, p. 49.) The set of assets heretofore considered include liquid financial assets. Friedman and Meiselman discovered that the dual criteria were best satisfied by the sum of currency and all privately held deposits at commercial banks (pp. 182–84).

Kaufman examined the proposition that if money is a factor in determining *GNP*, its effect may be delayed by as long as a year. From correlations between *GNP* and various financial aggregates which led *GNP* by +4 to −2 quarters, he found that the best definition of money depends on the number of quarters by which the financial measure leads or lags *GNP*. In general, the broader aggregates perform better when they are observed two or more quarters before income while the narrow definition performs best when observed concurrently with income (see Kaufman, pp. 86–87, Tables 1 and 2).

Kaufman examines the impact of changes in the monetary aggregate on changes in income in a particular current or future quarter, a procedure which is appropriate if money's effect on income occurs with a discrete time lag.

* Economist, Office of Management and Budget. This note has benefitted from discussions with J. L. Pierce and B. Friedman. The empirical calculations were largely performed by J. Walton. The author is solely responsible for the views expressed.

The present study extends the Kaufman analysis, allowing the effect of money on *GNP* to be distributed over several quarters by examining regressions of the following form:

$$GNP_t = a_0 + a_1M_t + a_2M_{t-1} + a_3M_{t-2} + a_4M_{t-3}$$

The distributed lag model examines the contention that the effect of changes in money spread slowly through the economy and that the long-run impact on *GNP* may build up over several quarters. Put differently, this approach investigates the proposition that knowledge of the time path of monetary changes might be more useful than a single quarter's change in determining the change in *GNP*.

I. The Correlation Tables

Data were collected in accordance with Kaufman's technique. Four end-of-month observations of various financial stocks were averaged to arrive at an average quarterly stock. Financial aggregates were then constructed : $M1$ = currency plus demand deposits; $M2 = M1$ plus time deposits at commercial banks; $M2.5 = M2$ plus deposits at mutual savings banks; $M3 = M2.5$ plus savings and loan shares; $M4 = M3$ plus savings bonds and postal savings deposits; and $M5 = M4$ plus private holdings of *U.S.* government short-term marketable securities. (A second method added currency last, following Kaufman's observation that currency is more closely related to income observed in earlier periods (p. 86), implying that income determines the demand for currency but currency is not a driving force determining income.) The financial variables and *GNP* were transformed to logarithms and expressed as first differences. The data here extend Kaufman's observations by two recent years; they cover the period 1953–1968. The correlations

Table 1—Aggregate Correlation Matrices

	Quarters Money Leads *GNP* +4	+3	+2	+1	0
1953–68					
Currency	−.01	.04	.14	.26	.40
*M*1 = Currency + *DD*	.04	.26	.41	.51*	.49*
*M*2 = *M*1 + *TD*	.28	.41*	.49*	.47	.32
*M*2.5 = *M*2 + *MSB*	.29	.41*	.49*	.47	.31
*M*3 = *M*2.5 + *SC*	.28	.41*	.48*	.45	.27
*M*4 = *MS* + *SB*	.31	.44*	.50*	.46	.30
*M*5 = *M*4 + *U.S.*	.03	.07	.19	.30	.39
1953–59					
Currency	−.61	−.59	−.30	−.14	.32
*M*1	−.40	−.15	.24	.63	.70*
*M*2	−.17	.03	.35*	.55	.30
*M*2.5	−.17	.04	.37*	.55	.29
*M*3	−.15	.07	.41*	.60	.33
*M*4	−.15	.10	.46*	.64	.36
*M*5	−.37	−.65	−.44	−.10	.30
1960–68					
Currency	.29	.32	.30	.39	.43
*M*1	.34*	.61*	.59	.47*	.32
*M*2	.51*	.66*	.62*	.44	.21
*M*2.5	.53	.67*	.62*	.44	.20
*M*3	.50	.66*	.57	.35	.07
*M*4	.54	.67*	.59	.37	.10
*M*5	.10	.48	.67*	.61*	.40

Source: Federal Reserve Board.

The variables are quarterly first differences of natural logarithms. *MSB* = deposit at mutual savings banks, *SC* = shares at savings and loan associations, *SB* = savings bonds and postal savings deposits, and *U.S.* = private nonbank holdings of *U.S.* government marketable securities maturing within one year.

* The entries marked with an asterisk satisfy the criterion that the aggregate correlation coefficient exceeds those of *GNP* with its components separately.

and regressions have been calculated for the whole period and the two subperiods, 1953–1959 and 1960–1968.

Correlation matrices shown in Table 1 were calculated as in Kaufman's study on the assumption of a point-input point-output relationship. Only the aggregate correlations are shown, but the asterisk indicates that the second criterion was satisfied. The results are essentially those found by Kaufman.

First: For the entire period the narrow definition of money (*M*1) performs best and satisfies the dual criteria when observed concurrently or one quarter before *GNP*; *M*4, which includes all of the "liquid assets" except marketable short-term government bonds, performs best when observed two or three quarters before *GNP*. In fact, the short-term government bonds, and *M*5 which includes them, do very poorly by these tests. (Adding currency last raises the correlations by about .02. This table is not shown.)

Second: The results for the two subperiods differ substantially from each other. The large current effect of *M*1 appears only in the earlier subperiod while the effect of money two or more quarters ahead of income is derived largely from the recent subperiod. The correlations with *U.S.* government bonds and *M*5 are strongly positive only in the recent period as are those of currency, while the correlations with *M*1 do not dominate those of the broader definitions in any lead quarter of that subperiod.

TABLE 2—F-STATISTICS TO TEST THE EQUALITY BETWEEN COEFFICIENTS OF MONEY IN REGRESSIONS OF THE FORM $GNP=a+bM_{t-i}$ (CHOW TEST) FOR THE PERIODS 1953–59 AND 1960–68

	t+4	t+3	t+2	t+1	t
Currency	14.55	14.27	4.39	1.78	.77
*M*1	7.36	4.11	1.05	6.21	8.72
*M*2	2.63	.66	.61	4.16	1.16
*M*2.5	2.61	.66	.88	4.15	1.12
*M*3	2.40	.59	1.50	5.94	2.18
*M*4	2.12	.19	2.21	7.65	2.18
*M*5	5.44	19.77	11.97	3.05	.52

The F-statistic with (3, 59) degrees of freedom is significant at 3.15 (5%) and 4.98 (1%).

II. Stability of the Relationships

To test Kaufman's conclusion that "These findings do not differ greatly for the two subperiods" (p. 84), with the extended data, the Chow test (see J. Johnston, pp. 136–38) was performed on simple regressions of *GNP* on money for the three time spans. The resulting F-statistics are presented in Table 2. Clearly, many of the coefficients have changed significantly. The improved performance of currency and *M*5 are the most marked, while the decline in impact of *M*1 in periods t and t+1 and its improvement in t+3 and t+4 are also highly significant. When the subperiods are compared to the correlations for the whole period it is clear that the large immediate impact of *M*1 is due mostly to the period before 1960 and that the performance of the broader aggregates is perhaps more relevant today. An examination of the recent subperiod casts doubt on the timeliness of Kaufman's contention that *M*1 satisfies the dual criteria best (p. 86).

A developing financial sophistication of the household and business sectors over the recent years may have led to some of the differences between the two subperiods. For instance, there are over twice as many asterisks in the recent subperiod than in the earlier one, indicating that asset substitution is becoming more widespread. The increasingly competitive nature of the financial markets, particularly in the 1960's when the demand for credit drove short-term interest rates to historic levels, has alerted the public to the desirability of allocating assets according to conditions of supply (relative interest rates).

One explanation for the dramatic change in the role of short-term government securities since 1960 is that corporate holdings of bills for tax anticipation purposes have been brought more into line with the timing of income earned.[1] Prior to 1960, tax payments lagged behind the realization of profits by as much as a year, and the bills which were accumulated for this purpose lagged the economic activity (*GNP*) which gave rise to the profits. Some support for this proposition is given by the large positive correlations in the 1953–1959 period between *GNP* and short-term bonds observed one and two quarters later (see Kaufman, p. 80). Since 1960, tax payments have been made gradually more synchronous with profits earned, and the bills which are held as tax hedges have tended to be accumulated more nearly when the level of business activity is highest. Since tax transactions are one expense of doing business, this source of demand for liquidity properly belongs here.

In addition, corporate treasurers have used an increasing portion of their holdings of bills in their general liquidity portfolio. The rapid development of the negotiable time certificate of deposit (*CD*'s) since 1961 has contributed to this development. *CD*s are well suited to perform the tax hedge role: a business can solidify a customer relationship while earning interest on money

[1] This idea was raised in discussions with B. Friedman.

which would not otherwise be free for other uses. Banks accommodate this need by bunching the *CD* maturities around tax dates. The behavior of the secondary *CD* interest rate attests to the similarity of *CD*s and other short-term marketable bonds.

$$RCD = \underset{(6.53)}{.574\ RTB} + \underset{(5.01)}{.432\ RCP} + \underset{(5.75)}{35.976\ (CD/GNP)} - \underset{(-5.78)}{.590}$$

$R^2 = .994$ $D.W. = 1.51$ Period: 1961–IV to 1968–IV $S.E. = .07$ $D.F. = 15$

In this quarterly regression, a *CD* demand function has been normalized on the secondary *CD* rate (*RCD*), and includes strong effects of the bill rate (*RTB*) and the commercial paper rate (*RCP*).[2] In this environment, including strong rate substitution and the increased availability of other assets which can perform the tax hedge role, bills are largely freed to perform the liquidity role, contributing to their increases in "moneyness" since 1960.

Further evidence which supports these tentative conclusions can be found in W. H. Locke Anderson's results. In his time period, 1948–60, corporate short-term bond holdings bear very little relationship to sales, his transactions variable, but are highly correlated with accrued tax liabilities as the present hypothesis implies. In some of his capital expenditure equations, the coefficients on short-term bonds and tax accruals are essentially equal in value but opposite in sign, indicating that only the excess of bonds over tax accruals is relevant for the real investment decision (see pp. 45–51, 75–76, especially equation (5)), another implication of the hypothesis suggested above. It would be interesting to examine these relationships for the period since 1960: the expectation is that they would no longer hold.

[2] The observations in which *RCD* exceeded the Regulation Q ceiling have been omitted since the change in outstandings largely represents a runoff from the previous maturities in those periods. This regression is due to H. T. Farr.

III. Distributed Lags

In an attempt to investigate further the implication of a lagged response of *GNP* to monetary aggregates and to try to achieve a more reliable prediction of *GNP*, distributed lags (rather than discrete lags) were estimated by regressing *GNP* against current and several prior observations on the monetary variables. In this framework, the total effect of money on *GNP* could be accumulated over several periods. The results of this experiment are displayed in Table 3.

The application of distributed lags does not change the conclusions drawn from the discrete lag experiment.

First: For the whole period and early subperiod, the narrow definition of money ($M1$) performs the best. Adding lags beyond $L=1$ to $M1$ does not improve its predictive power. In the 1953–59 period, the coefficient of determination is rather large for $M1$, but the application of distributed lags does not significantly improve on the correlations of Table 1. The sums of coefficients on lagged values of $M5$, which includes short-term securities, are sometimes negative, contrary to expectation.

Second: For the recent subperiod, the broader measures perform best. $M5$ has perhaps the best record: its estimation of *GNP* is about the best with only the current and two lagged observations. The large predictive power of $M1$ exhibited in the early subperiod has been eroded. These results essentially parallel those of the simple correlation table.

IV. The Role of an Empirical Definition of Money

The usefulness of the debate over the definition of money can be questioned on methodological grounds as well as on its relevancy to policy questions. The empirical method used to define money should correspond to the form of the money demand function in the underlying model. Leaving aside for the moment the well-known debate concerning the presence of absence of interest rates in this function (see Galper, Hamburger (1969), and Lee), there remains the issue of

TABLE 3—VALUES OF ADJUSTED R^2 FOR THE DISTRIBUTED LAG EQUATIONS

$$\Delta\ Ln\ GNP = a + \sum_{i=0}^{L} b_i \Delta\ Ln\ M_{t-i}$$

	$L=4$	$L=3$	$L=2$	$L=1$	$L=0$
1953–68					
Currency	.21	.19	.14	.15	.15
$M1$ = Currency + DD	.29	.28	.29	.29	.23*
$M2 = M1 + TD$	.21	.22	.22	.20	.09
$M2.5 = M2 + MSB$	.20	.21	.21	.20	.08
$M3 = M2.5 + SC$	.20	.21	.21	.18	.06
$M4 = M3 + SB$	.21	.23	.23	.20	.07
1953–59					
Currency	.38	.38	.21	.14	.07
$M1$	.48*	.51*	.52*	.53*	.47*
$M2$	.17	.19	.22	.25	.06
$M2.5$	.16	.19	.21	.24	.05
$M3$	.21	.26	.28	.30	.07
$M4$	.30	.33	.35	.37	.09
$M5$	.43[a]	.46[a]	.25[a]	.08	.06
1960–68					
Currency	.11	.11	.13	.15	.16
$M1$	.40	.42	.31	.18*	.07
$M2$	.42	.43	.33	.16*	.01
$M2.5$	.43	.44	.33	.17*	.02
$M3$	.40	.40	.27	.12	−.02
$M4$	.42	.43	.30	.15	−.02
$M5$	.46	.43	.44*	.34*	.14

* The asterisk signifies satisfaction of the dual criteria.

[a] These entries exhibit a negative sum of coefficients contrary to expectation.

the appropriate scale variable in the relationship.

If money is held mainly to facilitate transactions, the appropriate scale variable to use in the empirical definition of money should include financial and interindustry transactions as well as those associated with the purchase of final goods and services. If money demand is related to wealth in a portfolio balance model, wealth or permanent income would reflect the underlying assumptions more accurately than would current *GNP*. Thus, the wealth constraint would appear to be relevant to the more broadly defined monetary aggregates while the transactions motive applies more closely to the narrowly defined medium of exchange. In either case, *GNP* is not the best variable to use although it probably reflects transactions more closely than it reflects wealth. By this reasoning, the tests presented here and elsewhere in the literature may be biased against the more inclusive aggregates in the attempt to discover the appropriate definition of money.

The usefulness of the empirical definition of money from the policy point of view depends on the structure of the underlying model. In particular, the endogeneity of the money supply becomes particularly troublesome to policy makers if the best definition includes the liabilities of non-bank financial intermediaries. Given a model in which asset stocks are endogenous, policy multipliers can be obtained from either reduced form equations in which all exogenous variables are represented, or simulations of a structural model containing deposit supply as well as

demand relationships. Discussions of the empirical definition of money represent neither alternative adequately, and hence their usefulness to the policy making process is limited. (See Ando and Modigliani, De Prano and Mayer, Friedman and Meiselman, and, more recently, de Leeuw and Kalchbrenner.)

One endogenous financial variable which has not been considered in the present exercise is the liquidity represented in the value of policy reserves of life insurance companies. Policy loans made at 5 percent have become important as market rates have climbed above that value. A regression of reserves less policy loans as a ratio to net worth on various variables clearly establishes this asset as a substitute for savings deposits and marketable assets.[3]

$$(R - PL)/NW = - \underset{(-2.1)}{.14\,RCB} - \underset{(-3.9)}{.43\,ARSD} - \underset{(-3.4)}{.11\,ARCP} + Z$$

$$R^2 = .997 \quad DW = .96 \quad DF = 45$$

The equation indicates the large interest rate effect from savings deposits, $ARSD$, as well as substitution with corporate bonds, RCB, and commercial paper, $ARCP$. The inclusion of endogenously supplied substitutes for financial assets affects the solution of the system and the magnitude of the reduced form coefficients. This "asset," which has been ignored until recently (see Hulett), should be considered in specifying a full financial model.

In summary, the Friedman-Meiselman-Kaufman exercise of empirical money definition may be questioned. Its results are not particularly stable, nor do they "explain" a large fraction of the variance in *GNP*. *GNP* is not the best scale variable, and its use may bias the results in favor of narrow definitions of money. Also, its usefulness for the exercise of policy founders on the two problems of endogeneity of certain included variables and omission of various exogenous variables needed for the reduced form model (see Laidler). If one wants to study asset substitutability, perhaps the best way is the frontal assault which has been attempted by various authors including Chetty, Gramlich and Hulett, and Hamburger.

[3] In this equation (from Gramlich and Hulett) $R-PL$ is life insurance company reserves less policy loans; NW is net worth of the public; RCB is the corporate bond rate; $ARSD$ is a three period Almon distributed lag on a weighted average of the rates on savings deposits and consumer *CD*s at commercial banks, deposits at mutual savings banks, and savings and loan shares; $ARCP$ is a distributed lag on the commercial paper rate; and Z represents terms in disposable income, capital gains, and the constant term.

REFERENCES

W. H. L. Anderson, *Corporate Finance and Fixed Investment*, Boston 1964.

A. Ando and F. Modigliani, "Velocity and the Investment Multiplier," *Amer. Econ. Rev.*, Sept. 1965, *55*, 693–728.

V. K. Chetty, "On Measuring The Nearness of Near-Moneys," *Amer. Econ. Rev.*, June 1969, *59*, 270–81.

F. de Leeuw and J. Kalchbrenner, "Monetary and Fiscal Actions: A Test of Their Relative Importance in Economic Stabilization—Comment," *Review*, Federal Reserve Bank of St. Louis, Apr. 1969, 6–11.

M. DePrano and T. Mayer, "Autonomous Expenditures and Money," *Amer. Econ. Rev.*, Sept. 1965, *55*, 729–52.

M. Friedman and D. Meiselman, "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States, 1897–1958," in E. C. Brown et al., eds., *Stabilization Policies: Commission on Money and Credit*, Englewood Cliffs 1963.

M. Friedman and A. J. Schwartz, "Monetary Statistics of the United States," unpublished manuscript, Nat. Bur. Econ. Res.

H. Galper, "Alternative Interest Rates and The Demand For Money: Comment," *Amer. Econ. Rev.*, June 1969, *59*, 401–07.

E. M. Gramlich and D. T. Hulett, "Savings Flows, Mortgage Markets, and Residential Construction in the FRB-MIT Econometric Model," *Conference on Savings and Residential Financing*, U.S. Savings and Loan League, Oct. 1970.

M. Hamburger, "The Demand for Money by Households, Money Substitutes, and Mone-

tary Policy," *J. Polit. Econ.*, Dec. 1966, *74*, 600–23.

———, "Alternative Interest Rates and The Demand for Money: Comment," *Amer. Econ. Rev.*, June 1969, *59*, 407–12.

J. R. Hicks, *Value and Capital*, 2nd ed., London 1957.

D. Hulett, "The Mortgage Market: Alternative Specifications," mimeo. 1969.

J. Johnston, *Econometric Methods*, New York 1963.

G. Kaufman, "More on An Empirical Definition of Money," *Amer. Econ. Rev.*, Mar. 1969, *59*, 78–87.

D. Laidler, "The Definition of Money: Theoretical and Empirical Underpinnings," *J. Money, Credit, Banking*, Aug. 1969, *1*, 508–25.

T. H. Lee, "Alternative Interest Rates and the Demand for Money: The Empirical Evidence," *Amer. Econ. Rev.*, Dec. 1967, *57*, 1168–81.

Dependency Rates and Savings Rates: Comment

By Kanhaya L. Gupta*

In a recent issue of this *Review*, Nathaniel Leff examined the role of demographic factors in the determination of aggregate savings rates, using international cross-section data. His major conclusion was "... that dependency ratios are a statistically distinct and quantitatively important influence on aggregate savings ratios, both for the 74 countries considered as a whole and within the subsets of developed and underdeveloped countries" (pp. 893–94). In this note, we shall present evidence showing that, contrary to Leff's findings, dependency ratios play an insignificant role in determining savings rates in the majority of the underdeveloped countries.

In a recent paper, I argued and showed that the treatment of underdeveloped countries as a single group is not very meaningful. In fact, very often such a treatment conceals more information than it reveals. It was then argued that a more satisfactory way is to subdivide these countries according to per capita income levels. Following the classification adopted in that paper, I divided the underdeveloped countries into three groups: (I) those with per capita income between \$0–124; (II) those between \$125–249; and (III) those between \$250–675.[1] Table 1 gives the identification of the countries in each group. Using Leff's data, we estimated his equations for each group separately and for the three groups combined together. The equations are:

$$(1)\quad S/Y = f_1(Y/N, g, D_1, D_2)$$

$$(2)\quad S/Y = f_2(Y/N, g, D_3)$$

$$(3)\quad S/N = f_3(Y/N, g, D_1, D_2)$$

$$(4)\quad S/N = f_4(Y/N, g, D_3)$$

Following Leff, all the equations were estimated in log-linear form. The symbols used stand for:

S/Y = aggregate savings ratio
S/N = per capita aggregate savings
Y/N = per capita income
g = rate of growth of per capita income
D_1 = percentage of population aged fourteen or less
D_2 = percentage of population aged sixty-five or more
D_3 = total dependency ratio, the sum of D_1 and D_2

TABLE 1—COUNTRIES IN EACH GROUP

Group I (9 Countries)	
India	Taiwan
Kenya	Tanganyika
Pakistan	Thailand
South Korea	Uganda
Sudan	
Group II (16 Countries)	
Ceylon	Jordan
Dominican Republic	Mauritius
Ecuador	Morocco
Egypt	Paraguay
Ghana	Peru
Honduras	Philippines
Iran	Portugal
Iraq	Tunisia
Group III (22 Countries)	
Argentina	Japan
Barbados	Malaysia
Brazil	Malta
British Guiana	Mexico
Bulgaria	Nicaragua
Chile	Panama
Costa Rica	Poland
Cyprus	Spain
El Salvador	Trinidad-Tobago
Greece	Turkey
Jamaica	Uruguay

In Table 2, we present the results of equations (1) and (3) only, because they are really the equations capturing the effects of

* Associate professor, University of Alberta. I am grateful to Leff for providing me with the data.

[1] The results of an alternative classification adopted in the United Nations Report were similar.

Table 2—Results for Different Groups

Dependent Variable	Intercept	*Ln Y/N*	*Ln g*	*Ln* D_1	*Ln* D_2	$\bar{R}^2$	*F*
Group I							
(1) *Ln S/Y*	3.9549	0.4548 (0.3768)	0.1263 (0.3737)	−0.7685 (0.1481)	−0.6475 (0.9446)	−0.29	0.55
(2) *Ln S/N*	−0.3564	1.6112 (1.6008)*	0.1309 (0.4644)	−1.0487 (0.2424)	−0.5585 (0.9770)	0.403	2.35
Group II							
(3) *Ln S/Y*	5.7068	−0.1372 (0.4052)	0.0234 (2.437)**	−0.6172 (0.7274)	−0.0449 (0.1419)	0.240	2.19
(4) *Ln S/N*	2.7274	0.8330 (2.6194)**	0.0243 (2.6937)**	−0.9724 (1.2200)	−0.1404 (0.4726)	0.495	4.67
Group III							
(5) *Ln S/Y*	14.2616	−0.0584 (0.2401)	0.0347 (2.2646)**	−2.6974 (4.4973)***	−0.8866 (2.5979)***	0.564	7.78
(6) *Ln S/N*	8.9245	0.9172 (3.7661)***	0.0345 (2.2460)**	−2.4937 (4.1514)***	−0.7803 (2.2831)**	0.766	18.15
All Groups							
(7) *Ln S/Y*	9.3209	0.1624 (1.9444)**	0.0258 (3.1198)***	−1.8402 (4.5442)***	−0.5416 (2.700)***	0.471	11.22
(8) *Ln S/N*	4.6341	1.1501 (14.7254)***	0.0271 (3.5020)***	−1.8012 (4.7578)***	−0.5014 (2.7803)***	0.900	104.57

Notes: *t* values are given in the parentheses; * significant at 10% level; ** significant at 5% level; and *** significant at 1% level.

the demographic factors studied by Leff.[2] Looking at this table, we can draw the following inferences:

a) For Group I, the only significant variable is the level of per capita income. Furthermore, in terms of equation (2), even the combined effect of *ln* D_1 and *ln* D_2 fall short of the coefficient of *ln Y/N*. This is just the opposite of the results obtained by Leff for the underdeveloped countries.

b) For Group II, the results are slightly different in that now, in addition to the level of per capita income, the rate of growth of income is also significant. However, in so far as the effect of the dependency ratios is concerned, it still continues to be negligible.

c) For Group III, we get excellent results. Here all four variables are highly significant and their coefficients have the right signs. The effect of the dependency ratios is more important than that of per capita income. All these results are in full accord with Leff's findings. Looking at equations (7) and (8) in Table 2, we can observe that the results of Group III, are very similar to the results for the three groups combined together.

[2] The results of equations (2) and (4) are available from the author. The results of equation (3) test directly the hypothesis that the income elasticity of savings is significantly different from unity for the different groups. Since Leff also presents the results of this equation, it facilitates comparison of our results with his. We recognize that the $\bar{R}^2$s for equation (3) will be biased, as pointed out by the referee and Leff, because of the presence of Y/N on both sides of equation (3). This follows from the fact that $S/N = (S/Y)(Y/N)$.

d) The income elasticity of savings is not significantly different from unity for any of the groups.

It is thus clear that the overall results for the underdeveloped countries merely reflect the savings behavior of countries in Group III. Dependency ratios do not appear to play any role in the other two groups. Hence the failure of the aggregate savings ratio to rise in these countries will have to be explained by factors other than the demographic factors considered by Leff. It should be noted here that Groups I and II account for more than 50 percent of the countries included in the underdeveloped category.

The above findings do not mean that Leff's hypothesis is completely wrong. It is possible that his hypothesis represents an *inter-group* relationship and not necessarily an *intra-group* relationshipfor all the groups.[3] While this interpretation leaves something of Leff's hypothesis intact, still it sheds very little light on the failure of the savings ratio to rise in the countries of Groups I and II.

A possible explanation of the failure of dependency ratios in the equations for Groups I and II is as follows. When income levels are as low as in these two groups, there is no margin left for savings. In fact, even in the absence of dependents these income levels would not provide the minimum level of living. Hence, it simply means that people are only sharing poverty—having two meals a day can lead to no more savings than only one meal a day. We suggest that demographic factors, like the dependency ratios, become operative and significant only when the per capita income of the *working population* reaches a level where it can provide more than a minimum level of living, thus generating potential savings. These potential savings can then be prevented from being realized by the presence of dependents.

[3] The managing editor suggested this possibility and I am grateful to him for his suggestion.

REFERENCES

K. L. Gupta, "Foreign Capital and Domestic Savings: A Test of Haavelmo's Hypothesis with Cross-Country Data: A Comment," *Rev. Econ. Statist.*, May 1970, *52*, 214–15.

N. H. Leff, "Dependency Rates and Savings Rates," *Amer. Econ. Rev.*, Dec. 1969, *59*, 886–96.

United Nations, *Report on the World Social Situation*, New York 1961.

Dependency Rates and Savings Rates: Comment

By NASSAU A. ADAMS*

In a recent paper in this *Review*, Nathaniel Leff concludes, on the basis of an analysis of cross-section data covering some seventy-four developed and underdeveloped countries, that demographic conditions are ". . . a major determinant of aggregate savings rates," that dependency ratios are ". . . a statistically distinct and quantitatively important influence on aggregate savings ratios, . . . " and that "High dependency ratios—and ultimately high birth rates—are among the important factors which account for the great disparity in aggregate savings rates between developed and underdeveloped countries" (pp. 893–94). In view of the central role which the aggregate savings rate has traditionally played in development theory, and in view furthermore of the growing concern in the international community with the question of population growth, the above conclusions linking these two issues clearly warrant attention.

The present note will examine the basis for Leff's conclusions at two levels. First, in terms of his theoretical rationale for expecting such a relationship; second, in terms of an analysis of his statistical results.

I

So far as theoretical rationale is concerned, the point is made that population growth typically means high birth rates and hence a high proportion of children in the population. And since "[children] . . . contribute to consumption but not to production, . . . [they] might be expected to impose a constraint on a society's potential for savings," (Leff, p. 887).

There seems to be two main weaknesses with this argument. First, it assumes that the factors that determine output are independent of the number of dependents for which provision has to be made. This need not be the case. It may well be that the pressure of increased family size on individual motivation and response favorably affects productivity and output. As Richard Easterlin puts it in a sobering discussion of the population issue, "Population pressure arising from mortality-reduction may provide the spur to work harder, search out information, increase capital formation, and try new methods" p. 104. This being so, the effect of higher population growth rates[1] may well be higher levels of output,[2] and perhaps higher levels of savings and capital formation as well to provide for the future needs of dependents, and at the abstract level of theorizing, it is not at all obvious that a negative relationship between these variables is to be expected.

The second weakness concerns the use of age fourteen as the cut-off point for dependent children. In the developed countries, or perhaps even in the cities of developing countries, this figure may well be appropriate, but it is totally inadequate in respect to farm families in the developing countries, most of whose children start to become economically useful at age six or seven. And since farm families account for the bulk of the population of the developing countries, the meaning of the analysis for these countries is clearly in doubt.[3]

* Economist, United Nations Conference on Trade and Development, New York. The views expressed are those of the author only. I am indebted to V. K. Sastry for helpful discussion of some of the issues raised in this note, as well as for the reference given in fn. 6.

[1] Resulting, e.g., from mortality reductions accompanied by unchanged high birth rates.

[2] And it may be noted here that among developing countries there is apparently a significant positive relationship between the rate of growth of output and the rate of population growth (see the article by John Conlisk and Donald Huddle).

[3] It would clearly be absurd to say that a farm child of ten years of age contributes to consumption but not to production. And since this proposition is the main basis for giving economic interpretation to the demographic variable (at least that involving child dependency), this is an important weakness in the analysis.

II

The equations Leff fitted to data for a single year, 1964, were of the form:

$$(1) \qquad S/Y = f(Y/N, g, D_1, D_2)$$

where

S/Y = the aggregate savings ratio
Y/N = per capita income
g = rate of growth of per capita income (annual average for the preceding five years)
D_1 = the proportion of population age fourteen and below
D_2 = the proportion of the population age sixty-five and above.

The rationale for introducing D_2 is similar to that for D_1, but it must be noted here that this variable has no bearing on the issue of population growth, high birth rate, etc., which is much emphasized by Leff in his conclusions.

These equations were fitted to data for seventy-four developed and developing countries grouped together and for forty-seven developing countries and for twenty western developed countries separately. Leff's results are summarized in Table 1. The dependency variables D_1 and D_2 appear significant in both the equation for all seventy-four countries as well as in that for the forty-seven developing countries, although in the latter the proportion of variation explained amounts to only 24 percent (compared with 57 percent for the all-country sample). These results provide the basis for Leff's conclusions.

There are several problems involved in Leff's interpretation of these results. The first point to note is that for the forty-seven developing countries, the dependency ratio D_1 varies only between 40–46 percent (and D_2 between 3–5 percent). It seems hardly convincing to argue that variations in the "dependency" ratio within a range of 40 and 46 percent could be a major factor accounting for differences in aggregate savings rate. But the important point here is that Leff presents no argument why the functional form

$$S/Y = f(Y/N, g, D_1, D_2),$$

is the theoretically correct one. In fact his arguments lead to the expectation that $S/Y = f(Y/N, g, D_1)$ should be just as valid, and indeed D_2 is mentioned almost as an afterthought in the development of the argument. The fact is, however, that using his own data for the forty-seven developing countries,[4] and fitting the equation $S/Y = f(Y/N, g, D_1)$, shows D_1 to be totally without statistical significance as an explanatory variable. The relevant results are summarized in Table 2. It appears from these results that when D_1 is introduced as the only dependency variable, it is without significance.

[4] I am grateful to Leff for kindly making available to me his data on which these results are based.

TABLE 1—REGRESSION COEFFICIENTS: ($\ln S/Y$ DEPENDENT VARIABLE)

	Independent Variables				Const	$\bar{R}^2$	F
	$\ln Y/N$	$\ln g$	$\ln D_1$	$\ln D_2$			
A	.1596 (2.8776)	.0254 (3.2792)	−1.3520 (4.6406)	−.3990 (2.5623)	7.3439 (5.7289)	.5697	25.1604
B	.1292 (1.8487)	.0227 (2.8079)	−1.2297 (2.7636)	−.4455 (2.1554)		.2419	4.6685
C	.0035 (.0296)	.2589 (1.6228)	−.4324 (1.7099)	−.4916 (2.6547)		.4395	4.7245

Note: *t*-ratios in brackets.
A: All 74 countries; B: 47 developing countries; C: 20 western developed countries.

TABLE 2—REGRESSION COEFFICIENTS, 47 DEVELOPING COUNTRIES

	Independent Variables					Const.	$\bar{R}^2$
ln S/Y	*ln Y/N*	*ln g*	*ln D_1*	*ln D_2*	*ln D_3*		
(1)	0.0687 (0.97)	0.0771 (1.81)	−0.45921 (−1.40)			3.8889 (2.74)	0.1348 (0.13)
(2)	0.1251 (1.64)	0.0894 (1.97)		−0.668 (−0.44)		1.9537 (5.37)	0.0993 (0.10)
(3)	0.0678 (0.98)	0.0798 (1.90)			−0.8036 (−1.64)	5.2530 (2.58)	0.1485 (0.15)
(4)	0.1230 (1.75)	0.1134 (2.66)	−1.3753 (−2.87)	−0.5462 (−2.50)		7.7041 (3.79)	0.2289 (0.23)

Similarly for D_2, as well as for D_3 (defined as D_1+D_2). However when D_1 and D_2 are introduced together as separate variables they become highly significant. But these results seem to be a mere statistical curiosity, since the data are evidently plagued by multicollinearity, D_1 and D_2 being highly correlated, as shown in the correlation matrix of Table 3.

TABLE 3—CORRELATION MATRIX, BASED ON LEFF'S DATA FOR 47 DEVELOPING COUNTRIES

(all data in logarithmic form)

	Y/N	g	D_1	D_2	D_3
S/Y	.28	.32	−.32	.17	−.33
Y/N		.19	−.42	.52	−.37
g			−.17	.35	−.12
D_1				−.81	.98
D_2					−.70

Multicollinearity is also indicated by the low proportion of variation explained by equation (4), i.e., 23 percent, compared with the high proportion of D_1 explained by D_2 and vice versa, i.e., 64 percent.[5] The instability of the coefficients, i.e., the fact that the coefficients for D_1 and D_2 change greatly when they are introduced together as against separately (compare equations (1) and (2) with equation (4)) only emphasizes further the problem of multicollinearity in the data.

And since in the presence of multicollinearity the standard error of the regression coefficients might be quite misleading,[6] the fact that both D_1 and D_2 appear with low standard errors in equation (4) cannot be accepted as establishing the statistical significance of these variables.

It may therefore be concluded that the results shown in Leff's paper are statistically unreliable, and that it is therefore quite out of the question to conclude from them that dependency ratios, birth rates, population growth, etc., are important factors affecting the aggregate savings rate. More careful analysis of the data reveal that they permit no such conclusions to be drawn.[7]

III

The major conclusion reached by Leff, that "High dependency ratios—and ultimately high birth rates—are among the important factors which account for the great disparity in aggregate savings rates between developed and underdeveloped countries" can now be seen to be entirely unwarranted. The fact is that developed countries tend to have high savings rates and low birth rates, and that for the underdeveloped countries the situation tends to be the reverse. This

[5] Thus regression of D_1 against D_2 yields (in log form)

$$\ln D_1 = 4.1607 - 0.3322 \ln D_2 \qquad \bar{R}^2 = 0.64$$
$$(82.45) \quad (-9.11)$$

[6] For a discussion of the problem of interpreting standard errors in the context of multicollinearity, see the article by Richard Stone.

[7] It is in any event difficult to see how such strong conclusions could have been drawn from an equation which was only able to explain 23 percent of intercountry variations in the savings rate.

being the case, by grouping developing and developed countries together, birth rates (or the corresponding dependency ratios), will "explain" differences in savings rates between these groups of countries. In fact a dummy variable taking the value 1 for developed countries and 0 for developing countries will serve equally well, as the following results show.[8] (Z is the dummy variable):

$$\begin{aligned} \ln S/Y = {} & 2.2826 + 0.0481 \ln Y/N \\ & (7.717) \quad (.8961) \\ & + 0.0986 \ln g + 0.5019\, Z \qquad (2) \\ & (2.6455) \quad (4.3962) \\ & \bar{R}^2 = 0.56 \end{aligned}$$

The results here closely parallel Leff's equation (3); the proportion of variation explained is approximately the same (56 percent), and the role played by the dummy variable corresponds to that of the dependency variable in his equation, the level of significance being approximately the same for both variables.[9]

From these results it is hardly possible to argue that the birth rate is an important factor affecting differences in savings rates. Clearly a host of factors determine birth rates, and these are evidently highly correlated with the level of development, per capita income, etc. A host of factors also determine savings rates, and these are also evidently correlated with the level of development, per capita income, etc. To group developed and developing countries together and show an inverse relationship between savings rates and birth rates hardly provides a basis on which to draw conclusion about economic phenomenon. And as we have seen, the relationship breaks down when the analysis is confined to the developing countries. And this does not seem to be merely a statistical phenomenon. As was argued earlier, any economic relationship between birth rates (or child dependency ratio) and the aggregate savings rate can only be a very tenuous one, and it would seem rather far-fetched to suppose that amidst the welter of factors that undoubtedly affect intercountry difference in aggregate savings, this could be strong enough to be statistically significant.

[8] Based on Leff's data for seventy-four developed and underdeveloped countries.

[9] These results follow from the fact that developed and underdeveloped countries form two clusters, the dependency ratio (D_1) ranging between 24-31 percent for the former and for the latter between 40-46 percent.

REFERENCES

J. Conlisk and D. Huddle, "Allocating Foreign Aid: An Appraisal of a Self-Help Model," *J. Develop. Stud.*, July 1969, *6*, 245–51.

R. A. Easterlin, "The Effects of Population Growth on the Economic Development of Developing Countries," in R. D. Lambert, ed., *The Annals of the American Academy of Political and Social Science*, Philadelphia 1967.

N. Leff, "Dependency Rates and Savings Rates," *Amer. Econ. Rev.*, Dec. 1969, *59*, 886–96.

R. Stone, "The Analysis of Market Demand," *J. Roy. Statist. Soc.*, 1945, *108*, pt. 3–4, 286–382.

Dependency Rates and Savings Rates: Reply

By NATHANIEL H. LEFF*

My paper presented data on two questions relating to the determinants of international savings rates. First, it showed that both in developed and in less developed countries, cross-section analysis did not indicate the level of per capita income to be an important determinant of S/Y. Second, it showed that both for the overall sample of seventy-four countries and within the subsets of more developed and less developed countries, dependency rates are a significant factor influencing aggregate savings rates. In their comments, Kanhaya Gupta and Nassau Adams raise a number of questions concerning these results, especially as they apply to the less developed countries. I will reply to their comments in sequence.

I

Gupta has grouped the observations for the less developed countries according to their level of per capita income, and estimated equations with $ln\ S/Y$ as the dependent variable.[1] His estimates indicate that for the two lowest income groups of countries, neither the level of per capita income nor the dependency rates is a significant determinant of aggregate savings rates. I doubt that this result is due to the explanation which Gupta proposes: that at low income levels, "no margin is left for savings." W. A. Lewis long ago pointed out some of the fallacies of this argument (p. 346 *ff.*). Indeed, one might expect that at low income levels, both the precautionary and the future-income motives for saving and investment would apply with special force.[2] Rather, I suspect that the results presented by Gupta for the countries of Groups I and II may reflect the estimation bias introduced in these small samples by errors in measurement, both in the dependent and independent variables. Such errors might be especially severe in the poorest countries because of the difficulties of estimating the income and savings generated in agriculture, which are proportionately larger in these countries.

In any case, the crux of Gupta's comment is that the overall results for the underdeveloped countries, which indicate the importance of the dependency variables, "merely reflects the savings behavior of countries in Group III." An obvious way to establish this proposition rigorously would be a test of covariance such as that proposed by Gregory Chow. Gupta presents no such results. However, I have computed the F-statistics to test for the equality of the regression coefficients as between the estimates for the forty-seven country set and the estimates for the countries of Groups I, II, and III. The F-ratios are, respectively: 0.72, 0.53, and 0.45. Since the critical value of F (5, 37) is 2.47 at the .05 level, we cannot reject the null hypothesis that the subsets are part of the same regression structure as the entire sample.[3] Thus, contrary to Gupta's assertion, these countries belong statistically to the same regression structure as the estimates for the entire set of less developed countries, and the overall estimates apply also to them.

* Associate professor in the Graduate School of Business, Columbia University. I have benefited from discussions with Jacob Merriweather, M. Ishaq Nadiri, and Maurice Wilkinson in the preparation of this reply. They bear no responsibility for its contents.

[1] I confine my discussion to the results which Gupta presents for equations with S/Y as the dependent variable because the equations with S/N as the dependent variable are derived from the S/Y equations, and contain no independent behavioral information. As noted in my paper, $S/N = S/Y \cdot Y/N$. Consequently, given Y/N, S/N is determined by S/Y.

[2] It is also not at all certain that rates of time preference for countries vary inversely with levels of per capita income. See the discussion in Leff (pp. 618-20).

[3] The F-statistics cited in the text were computed from equations in which $ln\ Y/N$, $ln\ g$, $ln\ D_1$ and $ln\ D_2$ were specified as the independent variables. If $ln\ D_3$ is substituted for the other dependency variables, the results of the Chow test are similar. The F-statistics for the three groups are respectively: 1.10, 0.22, and 0.29. Similar results of statistically insignificant F ratios were obtained when the Chow test was applied to equations estimated with $ln\ S/N$ as the dependent variable.

II

Adams raises a number of other questions, both analytical and statistical. He states that it "seems hardly convincing to argue that variations in the dependency ratio within a range of 40 and 46 percent [in the less developed countries] could be a major factor accounting for differences in aggregate savings rates." That, however, is something to be settled after considering the data rather than asserted *ex cathedra*. He also suggests that I presented no theoretical justification for estimating equations in which both D_1, and D_2 are specified. The discussion on pages 887–89 of my paper, however, explains why both of these variables are theoretically relevant. Furthermore, as I will show below, his estimates of equations omitting one or the other of the dependency variables are biased by a serious error in specification.

Before discussing the statistical issues, however, let us consider Adams' analytical points. He suggests that the pressures of increased family size on individual motivation may raise productivity and output in less developed countries. This argument suffers from two weaknesses. First, it seems to imply that in the absence of larger family size, families will persist in sub-optimizing household allocational decisions. This goes against the abundant evidence which has accumulated concerning the economic rationality of micro-decision making in the less developed countries. One can, of course, argue the existence of household slack or leisure which, with larger family size, will be allocated to productive activities. We must distinguish, however, between the motivation and the capacity for such a response. In particular, as has often been noted, children are a time-intensive commodity (see Gary Becker). Apart from their simple consumption needs, the care of children involves large additional demands on household time and energies. Consequently, contrary to Adams' assertion, larger family size may have a debilitating effect, and *reduce* the capacity "to work harder, search out information, increase capital formation, and try new methods." At the least, higher dependency rates might be expected to lower the participation rate of women in the labor force.[4]

Adams also notes the difficulties involved in using age fourteen as the cut-off point for computing the dependency ratio D_1. To avoid some degree of arbitrariness in establishing a cut-off point, however, would require considerable knowledge concerning labor-force participation and worker productivity by age-cohort in the sample of seventy-four highly varied economies. In the absence of such data, I preferred to use a uniform procedure for all countries, and worked with the conventional demarcation point for the child dependency rate, age fourteen. Moreover, even if a fraction of the population aged fourteen or less does participate in the labor force in the less developed countries, their contribution to output is probably much less than the productivity of adults because of their inferior experience and strength.[5] It is striking that despite the obvious simplifications of the methodology and the rough nature of the data used, the effects of dependency rates on savings rates showed up so clearly in the equations estimated, both for the sample of seventy-four countries considered as a whole, and in the subsets of more developed and less developed countries.

III

Noting the high correlation between $ln\ D_1$ and $ln\ D_2$ for the set of forty-seven less developed countries, however, Adams suggests that because of multicollinearity, the results presented in my paper are a "mere statistical curiosity." We will examine below the possible bias introduced by multicol-

[4] The empirical evidence which Adams mentions from the study by John Conlisk and Donald Huddle, actually goes counter to his hypothesis. As discussed below, the correlation Adams cites indicates that in this sample, population growth was associated with a less than proportionate increase in the rate of growth of *per capita* output.

[5] Detailed data available for one less developed country, Mexico, do not indicate a high participation rate for the population aged fourteen or less. In 1960, the labor force participation rate for the cohort aged ten to fourteen years was only 15 percent. This relatively low rate obtained even though approximately 52 percent of the total labor force was in agriculture. See the study by El Colegio de Mexico (pp. 162, 175).

linearity. First, however, we should note that there is a straightforward way of clarifying this issue. In the equation where the total dependency burden, D_3 ($=D_1+D_2$), is specified, the possible problem of collinearity which Adams raises does not arise. Consequently the estimates for equations specifying $\ln D_3$ become of central importance in resolving the question of the effects of dependency conditions on savings rates.

My paper (p. 890) presented the results of such an estimate using the observations for all seventy-four countries. The regression coefficient for $\ln D_3$ was -1.5, with a t-ratio of 4.9 and a *beta* weight twice as high as those for the other variables specified. Adams estimated a similar equation for the sample of forty-seven less developed countries. However, the results which he reports for this equation (equation (3) of his second table) present some difficulties in interpretation. Perhaps because of an imprecise computer program (see J. W. Langley), his results differ considerably from the estimates of the same equation using the same data, which Gupta reports.[6] They also differ from the estimates which I obtained using three different computer programs.[7]

Adams reports a regression coefficient for $\ln D_3$ of $-.8036$, with a t-ratio of 1.64 and an $\bar{R}^2$ for the equation of .1485.[8] By contrast, estimating the same equation with the same data, Gupta reports a regression coefficient almost twice as large, -1.55, with a t-ratio of 3.71, and an $\bar{R}^2$ for the equation of .410. My results for the significance of the coefficient on $\ln D_3$, obtained with three different computer programs, also indicated statistical significance above the .05 level. Thus the estimates for the equation in which $\ln D_3$ was specified, and in which possible bias because of collinearity between $\ln D_1$ and $\ln D_2$ is not present, confirm the significance of the dependency burden as a factor accounting for the variance in the aggregate savings ratio in less developed countries.[9]

More generally, returning to the equations in which $\ln D_1$ and $\ln D_2$ were specified together, Adams seems to consider that the presence of a high simple correlation between two independent variables *ipso facto* results in valueless estimates because of multicollinearity. If multicollinearity biased the estimates of the equation in which $\ln D_1$ and $\ln D_2$ are specified together, however, we would expect that the estimates for these variables would lose their minimum variance properties, and the standard errors of these coefficients would increase. As seen in equation (4) of Adams' second table, however, precisely the opposite occurs: the t-ratios of the variables *rise* in the equation in which both $\ln D_1$ and $\ln D_2$ are specified together.

Finally, we should note that omission of the theoretically relevant, collinear variable *will* bias parameter estimates (Carl Christ, pp. 388–89, D. E. Farrar and R. F. Glauber, p. 106). This seems to be exactly what happened with Adams' equations in which he specified $\ln D_1$ without $\ln D_2$, and vice versa (equations (1) and (2) of his second table). This specification error (as well as the possible effects of imprecision in his computer program) may explain the relatively poor performance of Adams' equations in which $\ln D_1$ and $\ln D_2$ were employed separately.[10] Given the strong covariance between $\ln D_1$ and $\ln D_2$, it is not surprising that their coefficients change when one of these variables is omitted as compared with the equation when they are specified together.

[6] Gupta kindly made available his results for this equation.

[7] The computer programs which I used in estimating these equations were the NBER Stepwise Regression Program; MIAWOLS, written by M. Norman; and RAPE, for which the programming was done by W. J. Raduchel.

[8] A word on the critical t-values is in order here. Since we have definite theoretical expectations concerning the sign of the coefficient on the D variables, a one-tailed t test is appropriate. With 40 degrees of freedom, the critical values of t at the .05 and .1 significance levels are, respectively, 1.684 and 1.303.

[9] Adams may have been misled by an unreliable computer program. Still, he seems a bit cavalier in dismissing as "without significance" a coefficient which his own results indicate to have a t-ratio of 1.64, significant at a confidence level very close to .05.

[10] Adams indulges in something of an overstatement, however, when he says that when D_1 is specified without D_2, it is "totally without statistical significance as an explanatory variable." The t-ratio he presents (equation (1) of his second table) shows D_1 to be significant at the 10 percent level. As Kenneth Arrow has emphasized, dismissing results which are significant at this level may

IV

Adams also raises the issue of possible spurious correlation in the estimates relating to the sample for all seventy-four countries. That is, high dependency rates might be only a proxy variable for the general complex of conditions which comprise economic underdevelopment. I discussed this possibility in my paper (p. 891). To test for the possibility that the estimates showing the significance of the dependency variables were not simply picking up a "cluster effect," between the more developed and less developed countries, I estimated separate equations for the subsets of these countries. The results, presented in my paper, indicated the significance of the dependency variables within both subsets of countries.

In any case, the equation which Adams estimated with a dummy variable is hardly an adequate test for spurious correlation. A more appropriate way to test for the effects of dependency rates on savings rates—quite apart from the effects of the general complex of underdevelopment—would have been to specify the dependency variables *together* with the dummy variable to see if they retain their statistical significance. The results of an equation with this specification are presented below. A dummy variable was specified which took the value of 1 for the underdeveloped countries and 0 for the more developed countries. Student t-ratios are in parentheses, and the *beta* weights are underneath, in brackets.

$$(1)\quad \ln S/Y = \underset{(5.47)}{6.656} - \underset{(3.21)}{.398}\,\text{dummy} + \underset{\substack{(1.54)\\ [.205]}}{.083}\,\ln Y/N + \underset{\substack{(3.11)\\ [.233]}}{.023}\,\ln g - \underset{\substack{(3.26)\\ [.573]}}{.971}\,\ln D_1 - \underset{\substack{(2.70)\\ [.505]}}{.394}\,\ln D_2$$

$$\bar{R}^2 = .621 \qquad F \doteq 24.9$$

The t-value of the dummy variable is significant above the .01 level. This indicates that, not surprisingly, other features of underdevelopment aside from dependency conditions influence aggregate savings ratios. But the coefficients on $\ln D_1$ and $\ln D_2$ also retain their statistical significance above the .01 level. And while introduction of the dummy variable reduces the magnitude of the regression coefficient for $\ln D_1$ (and for $\ln Y/N$) as compared with the estimates (presented in my paper), for the equation without the dummy variable, the elasticities and the *beta* weights of the dependency variables continue to be very high.[11] Thus, contrary to Adams' assertion, equation (1) indicates that even when we control for other features of underdevelopment, dependency ratios are still important in accounting for the lower savings rates of the less developed countries.

V

Adams quotes a study showing a positive correlation between the rate of growth of output and the rate of population growth in developing countries. Adams might have noted, however, that the correlation he cites is between the rate of population increase and the rate of growth of *aggregate* rather than of per capita output (see Conlisk and Huddle, pp. 246–47). In fact, the elasticity of the aggregate growth rate with respect to

ad to Type I error and a costly loss in valuable information.

[11] Equation (1) was estimated with the observations for seventy-four countries, in order to permit comparison with Adams' results, which included all seventy-four countries. It is not clear, however, how to classify the observations for the seven Communist countries in this developed/underdeveloped dichotomy. Excluding the observations for those countries and reestimating equation (1) for the forty-seven less developed countries and the twenty developed countries gave the following results. The t-ratios are in parentheses and the *beta* weights are underneath, in brackets.

$$(1a)\quad \ln S/Y = \underset{(4.95)}{6.417} - \underset{(1.98)}{.270}\,\text{dummy} + \underset{\substack{(2.02)\\ [.411]}}{.118}\,\ln Y/N + \underset{\substack{(3.12)\\ [.183]}}{.023}\,\ln g - \underset{\substack{(3.20)\\ [.444]}}{.988}\,\ln D_1 - \underset{\substack{(2.46)\\ [.367]}}{.378}\,\ln D_2$$

$$\bar{R}^2 = .579 \qquad F = 19.2$$

This equation, which avoids the possible difficulties which may arise in the classification of the Communist countries, shows even more strikingly the effects of the dependency variable, even when the dummy variable is specified.

the rate of population increase in this sample was less than unity (0.69).[12] This result indicates that for this sample, higher population growth was associated with a *less* than proportional rate of increase in the growth of per capita *GNP*.

In any case, my paper related to savings rates rather than growth rates. The results implied that if it were not for the high dependency rates induced by high birth rates, less developed countries might be able to achieve higher savings ratios.[13] This might make possible higher rates of investment and growth than they would otherwise attain. In addition, they could finance a larger share of their investment with domestic savings rather than be dependent on foreign capital inflow, with its sometimes unwelcome side effects (see K. B. Griffin and J. L. Enos). Some views of development indeed stress precisely the capacity for autonomous economic progress, independent of the need for foreign resource flows. In this context, the term "dependency rates" takes on new, and broader meaning—dependency vis-à-vis the capital-exporting countries.

REFERENCES

K. J. Arrow, "Decision Theory and the Choice of a Level of Significance for the *t*-test," in I. Olkin et al., eds., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford 1960.

G. S. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

H. B. Chenery, "Targets for Development," Columbia University Conference on International Economic Development, New York 1970.

G. C. Chow, "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, July 1960, *28*, 591–605.

C. F. Christ, *Econometric Models and Methods*, New York 1966.

J. Conlisk and D. Huddle, "Allocating Foreign Aid: An Appraisal of a Self-Help Model" *J. Develop. Stud.* July 1969, *6*, 245–51.

L. E. Davis and R. E. Gallman, "The Share of Savings and Investment in Gross National Product during the 19th Century, United States of America," paper presented to the Fourth Congress of The International Economic History Association 1969.

D. E. Farrar and R. F. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," *Rev. Econ. Statist.*, Feb. 1967, *49*, 92–107.

K. B. Griffin and J. L. Enos, "Foreign Assistance: Objectives and Consequences," *Econ. Develop. Cult. Change*, Apr. 1970, *18*, 313 27.

A. C. Kelley, "Demographic Cycles and Economic Growth: The Long Swing Reconsidered," *J. Econ. Hist.*, Dec. 1969, *29*, 633-56.

K. A. Kennedy and B. R. Dowling, "The Determinants of Personal Savings in Ireland: An Econometric Inquiry," *Econ. Soc. Rev.*, Oct. 1970, *1*, 19–51.

J. W. Langley, "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User," *J. Amer. Statist. Ass.*, July 1967, *61*, 324–39.

N. H. Leff, "Marginal Savings Rates in the Development Process: The Brazilian Experience," *Econ. J.*, Sept. 1968, *78*, 610–23.

W. A. Lewis, *The Theory of Economic Growth*, Homewood 1955.

S. Robinson, "Aggregate Production Functions and Growth Models in Economic Development: A Cross-Section Study," unpublished doctoral dissertation, Harvard Univ. 1969.

T. P. Schultz, *A Family Planning Hypothesis and Some Empirical Evidence from Puerto Rico*, The RAND Corporation, Santa Monica 1967.

El Colegio de México, *Dinamica de la Población de México*, México 1970.

[12] The elasticities estimated in another study are somewhat lower, ranging from .35 to .55. (See S. Robinson as cited in Appendix Table 3 of Hollis Chenery.)

[13] My study was confined to a cross-section analysis. However, Allen Kelley's time-series country study of Australia (pp. 640–44) has also demonstrated an inverse relation between child dependency rates and personal saving. See also the results of a time-series study of personal savings in Ireland, by R. A. Kennedy and B. R. Dowling. Finally, in their discussion of the rise of the savings rate in the nineteenth-century United States, L. E. Davis and R. E. Gallman (pp. 21–22) also emphasize the importance of shifts in the dependency ratio.

A Property of a Closed Linear Model of Production: Note

By DAVID LEVHARI*

In a recent communication in this *Review*, Thomas Finn asserts that in a sufficiently well-behaved von Neumann model of production, the efficiency frontier tends to "flatten-out" as time goes on. He presents a graphic proof of this theorem. This proof is incomplete. The gap in the proof is the argument on page 534 of Finn's communication, ". . . the sequence of points a_1, a_2, . . . , cannot include a point that has coordinates greater than those of a^* (i.e., a_t cannot be farther than a^* from the origin) because we know that $\lambda^{(1-t)}F_t$ cannot be convex to the origin. Therefore we know that the distance between a_t and a^* converges to zero as t approaches infinity." The fact that a^* is an upper bound of the monotonic sequence of points a_t does not prove that a^* is the limit of this sequence. A monotonically increasing bounded sequence has the least upper bound as its limit but it has not been proved that a^* is, in fact, the least upper bound. I have tried to find a counter example that would fulfill all of Finn's assumptions and particularly his assumption that distinct points on the efficiency frontier F_2 must come from distinct input bundles. I was unable to find one that fulfills this assumption and thus it is still unknown to me whether the theorem stated is right or wrong.

In the following comment I would like to discuss the validity of the theorem for the general Closed Linear Model of Production. The flattening out property of successive transformation or efficiency frontier curves has been proved by E. Drandakis for the case of no joint production. The assumption "no joint production" is sufficient to assure that the asymptotic frontier is a hyperplane in any closed convex linear model in which the von Neumann ray is unique and the turnpike theorem holds. This is also discussed by J. R. Hicks in "The Story of a Mare's Nest" and there he poses the problem "to find out how much would be left of it (the flattening out) if the 'no joint supply' assumption were partially relaxed." An answer to this question is not given in the present comment and one still wonders whether non-joint production is also a necessary condition for the flattening to occur.

In the present note we support Hicks' assertion that the theorem does not hold for the general case of joint production and we show that there exists one class of well-behaved closed von Neumann models of production—interesting in their own right—where this property does not hold (however, they do not possess the property that distinct points on the efficiency frontier come from distinct input bundles). In this example the normalized efficiency frontiers in successive time periods coincide with one another. One wonders to what extent the example given below is exceptional. To answer this question, one would have to give a complete characterization of those dynamic models which do exhibit the property that their efficiency frontiers (in successive time periods) tend to hyperplanes. Here, a characterization will not be attempted.

Denote by S_t the vector of outputs produced at time t using the inputs S_{t-1} at time $t-1$.

The transformation of inputs into outputs is assumed to obey the following simple relation: $g(S_t)=f(S_{t-1})$ where g and f are homogeneous of the first degree and have the appropriate convexity-concavity properties (i.e., f is concave and g is convex). Both f and g are assumed differentiable. This is a completely well-behaved case.

The conditions for intertemporal efficiency are derived by maximizing one component of S_{t+1} given S_{t-1} and the rest of the components of S_{t+1}, that is by maximizing $g(S_{t+1})=f(S_t)$ subject to $g(S_t)=f(S_{t-1})$. Using Lagrange multipliers, we get that the condition for this is $f_{S_t^i}+\mu g_{S_t^i}=0$, $i=1$, . . . ,

* The Hebrew University of Jerusalem.

n where S_t^i is the ith component of S_t. By division:

$$\frac{f_{S_t^i}}{f_{S_t^j}} = \frac{g_{S_t^i}}{g_{S_t^j}} \qquad i, j = i, \ldots, n$$

This is the familiar condition that the marginal rate of substitution of inputs should equal their marginal rate of transformation as outputs. Using the homogeneity of f and g these equations are just sufficient to determine the unique ray of efficiency. As one expects, this ray is the von Neumann ray. In this case, we have the special feature that the economy is always on the turnpike (except in the initial and final stages).[1] Since we always start on the same ray, the shape of the normalized efficiency frontiers is always $g(S_t) = C$ where C is some constant depending on f for all t. For example, if the relation is

$$\sqrt{(S_t^1)^2 + (S_t^2)^2} = (S_{t-1}^1)^{\alpha} + (S_{t-1}^2)^{1-\alpha}$$

the shape of the efficiency frontier will be a quarter of a circle for all t.

It is still a puzzle to me whether Finn's theorem can be shown to be wrong also in other cases where the efficiency frontiers do not overlap.

[1] To see that this is really the Von Neumann ray' assume a balanced growth situation;

$$S_t = \lambda S_{t-1} \quad \text{or} \quad \lambda = \frac{f(S_t)}{g(s_t)}$$

Maximizing λ, we obtain:

$$\frac{f_{S_t^i}}{g_{S_t^i}} = \frac{f}{g} \qquad i = 1, \ldots, n$$

Using again the homogenity of the functions involved, we find that we have $n-1$ independent equations determing the ratios of the S_t^i or the ray in which:

$$\frac{f_{S_t^i}}{g_{S_t^i}} = \frac{f_{S_t^j}}{g_{S_t^i}}$$

This obviously describes the same rays as the equation in the text. Thus the condition of intertemporal efficiency implies that the economy is always on the von Neumann ray.

REFERENCES

E. Drandakis, *On Some Efficiency Properties of the Two Sectors Production Model*, Athens 1966.

T. J. Finn, "A Graphical Proof of a Property of a Closed Linear Model of Production," *Amer. Econ. Rev.*, June 1967, *57*, 531–35.

J. R. Hicks, "Prices and the Turnpike. The Story of a Mare's Nest," *Rev. Econ. Stud.*, Feb. 1961, *28*, 77–88.

Economics of Production from Natural Resources: Comment

By RICHARD F. FULLENBAUM, ERNEST W. CARLSON, AND FREDERICK W. BELL*

In a recent issue of this *Review*, Vernon L Smith advanced a general model of the economic aspects of production from natural resources. Unfortunately, Smith makes a number of unsound assumptions and methodological errors in formulating his general model of natural resource exploitation. We shall show, using commercial fishing as a particular example, that the structure of Smith's model leads to absurd conclusions. In addition, we shall develop the correct specification for the competitive recovery of a fishery resource.

I. Misspecification in the Smith Model

Smith attempts to develop his theory of production from natural resources by specifying a cost function. In the case of commercial fishing, Smith states that . . . "The most natural general hypothesis about total operating cost for the individual fisherman requires it to be an increasing function of the vessel's catch rate, x, but a decreasing function of fish population, i.e., $\phi=\phi(x, X)$, with $\phi_1\equiv\partial\phi/\partial x>0$, and $\phi_2\equiv\partial\phi/\partial X<0$" (p. 413). He then formulates the following cost function for the fishing firm,

$$C = \phi(x, X, K) + \pi \tag{1}$$

where C=total cost per unit of time; x =catch rate per vessel per unit of time; X=fish population; K=both the number of *homogeneous* firms (vessels) and a measure of the real capital stock in the industry; π =minimum return necessary to hold the vessel in the industry. As we shall see, the specification of (1) leads to unreasonable results. Actually, *total* cost per unit of time for any fishing firm is not a direct function

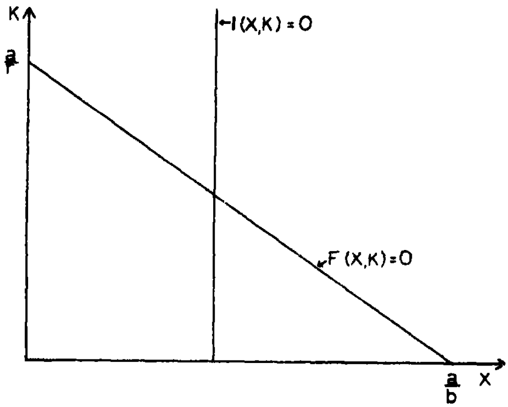


FIGURE 1. EXPLOITATION

of x, X, and K. Rather, total cost per firm per unit of time in long-run equilibrium is merely dependent upon factor prices. That is, total cost for the fishing firm *in the long run* is not explicitly influenced by output or x.[1] This is true since the output or catch rate per vessel, x, is determined by stock and technological externalities, e.g., $x=g(X, K)$. Assuming π represents the rate of return necessary to hold all factors in the fishing firm, the total cost *per unit of time* for the firm is simply equal to π. Total long-run *industry* cost is then equal to $K\pi$. Thus, average cost per unit of output, AC, for both firm and industry is given by,

$$AC \equiv \frac{K\pi}{Kx} \equiv \frac{\pi}{x} = \frac{\pi}{g(X, K)} \tag{2}$$

Smith has unknowingly introduced a tech-

* The authors are economists with the Division of Economic Research, Bureau of Commercial Fisheries, U.S. Department of the Interior. The comments in this article do not necessarily represent the official position of the U.S. Department of the Interior.

[1] For example, the northern lobsterman (or any other fisherman) will make the same expenditures per unit of time (i.e., fuel, bait, labor, interest, and insurance costs) regardless of the number of lobsters found in his traps. The number of lobsters caught depends on aggregate fishing effort which is beyond the control of the lobsterman. ***When we double the number of lobstermen exploiting a given fishery per unit of time, the total cost per unit of time is doubled, irrespective of what happens to catch rates.***

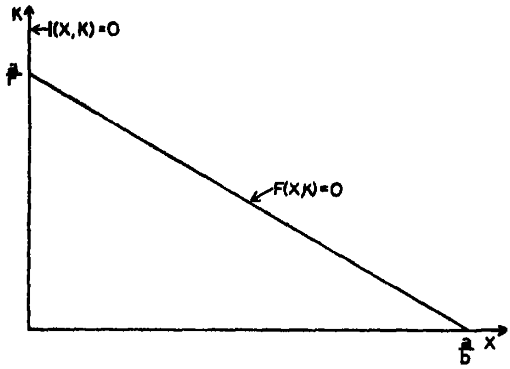


FIGURE 2. EXTINCTION

nological externality in the form of a market or pecuniary externality.[2] The basis of this misspecification stems from an apparent confusion between a short-run and long-run model of fishery exploitation. In the long run, any adjustments in the industry will be made in the number of *homogeneous operating units*, and not in the number of days fished per boat. However, a change in the number of units of time fished *does not change total cost per vessel* per unit of time fished. Secondly, if there is a variability in factor inputs per firm per unit of time—nowhere mentioned by Smith—then K is no longer a unique measure of the number of homogeneous vessels. His model must be corrected to eliminate these methodological errors since, as we shall see, they lead to untenable results.

[2] Put differently, our objection to Smith's model revolves about his cost specification and the manner in which he introduces the technological externality. In his system, the marginal cost of K (i.e., marginal cost of additional factors) is given by

$$(1') \quad \frac{\partial KC}{\partial K} = \frac{\partial}{\partial K}\left[K\phi(x, K, X) + K\hat{\pi}\right] = \frac{K\partial\phi}{\partial K} + \phi + \hat{\pi}$$

However, the properly specified marginal cost function for factors is

$$(2') \quad \frac{\partial KC}{\partial K} = \frac{\partial}{\partial K}(K\hat{\pi}) = \hat{\pi}$$

In (1′), the marginal cost of K depends upon K. However, the relationship between total cost and K is strictly linear and homogeneous, i.e., the opportunity cost of inputs is assumed unaffected by the expansion of the exploiting industry, and thus, the marginal cost of inputs is constant.

II. Competitive Recovery of a Fishery Resource

In this section, we shall present the classical system of commercial fishing behavior developed by Anthony Scott, H. Scott Gordon, J. A. Crutchfield and Arnold Zellner, and illustrate its usefulness by a quadratic example. At various points, comparisons will be made to show the erroneous nature of Smith's model. The general model (3)-(8) for the commercial fishing industry may be expressed as follows:[3]

$$(3) \quad \frac{dX}{dt} = f(X, Kx)$$

$$(4) \quad Kx = Kg(X, K)$$

$$(5) \quad KC = K\hat{\pi}$$

$$(6) \quad \pi = pKx - KC = pKg(X, K) - K\hat{\pi}$$

$$(7) \quad \frac{dK}{dt} = \begin{cases} \delta_1\pi, & \text{if } \pi \geq 0 \\ \delta_2\pi, & \text{if } \pi < 0 \end{cases}$$

where p = price, π = pure profit, δ_1 and δ_2 are the rates of entry and exit of capital, respectively. All other variables are defined above. In the above system, (3) is the biological growth function, (4) is the industry production function, (5) is the industry cost function, (6) is the industry profit function, and (7) is the industry entry function. The equilibrium condition for the industry ($\pi = 0$) may be formulated as follows:

$$(8) \quad p = \frac{R(Kx)}{Kx} = \frac{\hat{\pi}}{g(X, K)}$$

$R(Kx)$ is the revenue function for the industry. As shown in (4), the only way the variables X and K enter the profit function is through their respective impact upon the vessel catch rate, x.

Let us now specify a quadratic example of the classical system which assumes no crowding externalities, i.e., $\partial x/\partial K \equiv g_2 \equiv 0$. First we have a quadratic growth function or

[3] The more general model should include mesh considerations. Since Smith does not deal with this in his article, we shall abstract from mesh considerations in formulating our model.

$$\frac{dX}{dt} = aX - bX^2 \tag{9}$$

With exploitation of the fishery population by man, (9) becomes

$$\frac{dX}{dt} = aX - bX^2 - Kx \tag{10}$$

Thus (10) is a particular specification of (3). Under a steady state assumption (i.e., equilibrium between population growth and man-made mortality), $dX/dt=0$ and (10) becomes,

$$Kx = aX - bX^2 \tag{11}$$

Next, let us assume that the rate of loss to the population because of fishing is proportional to the number of firms (rK).[4] Multiplying the rate of loss because of fishing times the population, we have an industry production function or

$$Kx = rKX \tag{12}$$

Substituting (12) into (11) and solving for K in terms of X, we have

$$K = \frac{1}{r}(a - bX) \tag{13}$$

Therefore, the steady state relationship between K and X in (13) is inverse and linear. While linearity is not absolutely necessary, the following is, $dK/dX<0$.

The negative relationship between K and X is reasonable since increases in fishing effort or K will, *ceteris parabus*, always reduce the size of the stock. The inverse relation has been theoretically developed by M. Graham, M. B. Schaefer, T. I. Baranov, and R. J. H. Beverton and S. V. Holt and empirically verified by Schaefer, Beverton and Holt, and L. Van Meir. The inverse relation between X and K is one of the most fundamental laws of fishery population dynamics. Smith's Figure 3a is erroneous since it shows a positive relation for $F(X, K)=0$ between K and X below some value of X. *This is logical and biological nonsense.* Increasing the number of vessels, or K, cannot increase the size of the biomass *in any range.* Smith's error is directly traceable to the misspecification of the cost function and its use in deriving the relation between X and K.[5]

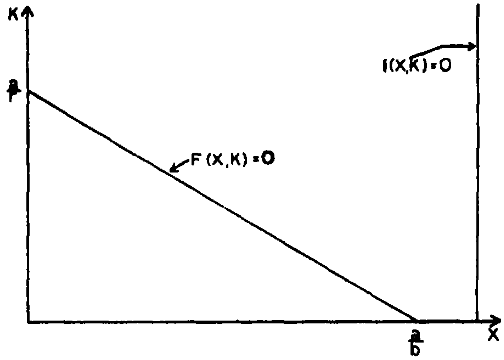


FIGURE 3. NON-EXPLOITATION

Assume a total revenue function (as does Smith) for the industry as follows (i.e., constant selling price),

$$R(Kx) = \bar{p}Kx \tag{14}$$

and a properly specified total cost function for the industry,

$$KC = K\hat{\pi} \tag{15}$$

or the number of vessels multiplied by their opportunity cost, $\hat{\pi}$. The profit function becomes

$$\pi = \bar{p}Kx - K\hat{\pi} \tag{16}$$

[4] This implies a linearly homogeneous production function, holding the biomass constant. However, the use of an industry production function, which reflects crowding externalities and thus decreasing returns with respect to the number of boats, e.g.,

$$Kx = rK^{\alpha}X, \qquad 0 < \alpha < 1 \tag{3'}$$

would not change any of our conclusions and would even be more realistic. Nonetheless, (12) was used because of its traditional role in the fishery literature.

[5] In particular, one could derive the curve $F(X,K)=0$ in the following manner. Assume a specified profit function for the fishing firm which has in it the type of cost function used by Smith. Taking the first partial with respect to x and setting equal to zero, one may derive an expression for x in terms of K and X, say $x=w(X,K)$. Then, under the assumption that $dX/dt=0$, we have $0=f(X)-Kw(X,K)$. The latter then is the implicit function $F(X,K)=0$, which may be solved in terms of K as a function of X.

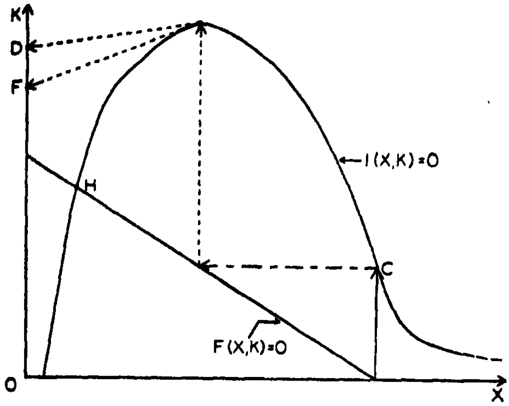


FIGURE 4

Competitive equilibrium is defined as a zero profit position or

$$(17) \qquad 0 = pKx - K\hat{\pi}$$

Since $Kx = rKX$, we may substitute in (17) and derive the equilibrium population with zero profit

$$(18) \qquad X = \frac{\hat{\pi}}{pr}$$

Equation (18) has the important property that the limit as p approaches ∞, $X \rightarrow 0$, and as price approaches 0, X approaches its natural upper limit. In (18) the population, X, increases with increases in the supply parameter $\hat{\pi}$ or fishing costs. When (18) is combined with (13), we can simulate 1) exploitation of the fishery resource, 2) nonexploitation, and 3) extinction as a lower limit to equation (18). Which of these three cases obtains depends upon the parameters p, $\hat{\pi}$, and r. These cases are shown in Figures 1–3.

The classical model can be adjusted to show extinction as a dynamic result. Assume that price is not parametric to the individual fishery, but rather that p is a linear function of Kx (quantity), so that total revenue is quadratic in Kx. Figure 4 shows one possible set of curves and possible dynamic path. If we start at C and follow the dashed line, extinction (point D or F) is possible, even though point H is a static equilibrium position.

III. Comparative Policy Implications

Because of the nature of Smith's cost function, he reaches some absurd conclusions in his discussion of sole ownership versus competition. Let us assume that price is constant to the individual fishery. Assume further that there are no crowding externalities, i.e., $C_3 \equiv 0$. Thus, Smith postulates the Lagrangian function for the sole owner as

$$(19) \qquad \psi = pKx - KC(x, X, K) + \lambda f(X, Kx)$$

Differentiating (19) with respect to K, X, and x, we find that

$$(20) \qquad p = C_1 + \lambda$$

where $\lambda = KC_2/f'$. If we want to obtain the sole ownership result, i.e., social marginal cost pricing, within a competitive free-access situation, a per unit of catch extraction fee, $U = \lambda$, should be imposed upon each fisherman. Under the assumptions of Smith's model $C_2 < 0$, and to the right of a population consistent with maximum sustainable yield, $f' < 0$. While Smith "proves" that under sole ownership, profit maximization guarantees exploitation of the biomass at the level at which $f' < 0$, the latter result does not necessarily follow in perfect competition. Suppose that with free exploitation, we are operating at X^*, the population at which total cost is equal to total revenue and that $\overline{X} < X^* < X^0$. It then follows that $f' > 0$, and $\lambda < 0$. At worst this implies a subsidy to a fishery which from an economic and biological point of view has been too intensively exploited.[6] At best, it demonstrates the lack of generality and lack of flexibility of Smith's "general" model.[7] On the other hand, the classical

[6] Overexploitation from a biological point of view usually refers to the reduction of the fishery population below that consistent with maximum sustainable yield.

[7] An anonymous referee has pointed out that 1) when $f' > 0$, Smith's tax does not apply because a sufficient condition for profit maximization for the sole owner requires $f' < 0$; and 2) that if $f' > 0$ in a purely competitive situation it is necessary to reduce fishing until the biomass is restored to its optimal level with $f' < 0$ and then to levy the tax mentioned by Smith in order to internalize social cost. Three comments seem appropriate. First, irrespective of the sign of f', a tax will reduce fishing effort! Thus, the imposition of a tax does not depend upon the second-order conditions for

theory can define a tax which depends upon the parameters but not the slope of the biological growth function. For example, assume sole ownership and the profit function as defined in (16). Maximizing (16) subject to the constraint that $dX/dt=0$, is identical to taking the inverse of (13), $K^{-1}(X)$, substituting in (12), finding Kx in terms of K alone, and then substituting back in (16). Thus, the profit function is specified solely in terms of K:

$$\pi = \frac{parK}{b} - \frac{pr^2K^2}{b} - K\hat{\pi} \tag{21}$$

Differentiating (21) with respect to K, setting equal to zero, we find

$$K^* = \frac{a}{2r} - \frac{\hat{\pi}b}{2pr^2} \tag{22}$$

K^* reflects the number of boats per unit of time consistent with marginal cost pricing.[8] Within the context of a purely competitive. free-access situation a tax per boat (T_κ) may be imposed so as to attain K^*:

$$T_\kappa = \frac{pK^*x^* - K^*\hat{\pi}}{K^*} \equiv px^* - \hat{\pi} = p\frac{ar}{2b} - \frac{\hat{\pi}}{2} \tag{24}$$

When (24) is added to $\hat{\pi}$ as an element of cost per firm, equation (21) becomes:

a profit maximum. Secondly, the restriction of fishing per se is a necessary and sufficient condition for a social optimum. If the optimal biomass, X^*, is attained through the restriction of fishing, then the optimal fishery catch K^*x^*, will also be reached and there is no need for a tax. Thirdly, within the structure of Smith's model *it is possible* that somewhere beyond MSY the amount of real capital exploiting a fishery is smaller under pure competition than under sole ownership. Then, the model requires an increase in the number of vessels in order to attain the sole ownership result; however, dynamically, this would move the biomass further away, not closer, to its optimal level! Thus, in any event, Smith's tax is meaningless.

[8] The solution in (22) may also be obtained by setting price (p) equal to marginal cost, so that:

$$p = MC = \frac{\hat{\pi}}{\frac{\partial Kx}{\partial K}} \tag{4'}$$

This yields the identical expression for K^*.

$$\pi' = \frac{parK}{b} - \frac{pr^2K^2}{b} - K(\hat{\pi} + T_\kappa) \tag{25}$$

Setting (25) equal to zero and solving for K we obtain the same solution as in (22). The tax per vessel does not depend upon the slope of the biological yield function; rather, it is positively related to the demand and supply parameters, p and r, respectively, and inversely dependent upon the opportunity cost per vessel. In addition, the tax, given the type of externalities assumed, will always be positive.

Thus, we have shown that the classical theory of commercial fishing is still on a firm foundation.

REFERENCES

T. I. Baranov, "On the Question of the Biological Basis of Fisheries," *Nauch. isseldov. Inst. Izv.*, 1918, *1*, No. 1, 81-128.

R. J. H. Beverton and S. V. Holt, "On the Dynamics of Exploited Fish Population," Ministry of Agriculture, Fisheries and Food, *Fishery Investigations*, 1957, Ser. II, *19*, 1–533.

J. A. Crutchfield and A. Zellner, Economic Aspects of the Pacific Halibut Fishery, *Fishery Industrial Research*, Apr. 1962, *1*, No. 1, Washington.

H. S. Gordon, "The Economic Theory of a Common-Property Resource: The Fishery," *J. Polit. Econ.*, Apr. 1954, *62*, 124–42.

M. Graham, "Modern Theory of Exploiting a Fishery and Application to North Sea Trawlers," *J. Conseil Int. Explor. Mer.*, July 1935, *10*, 264–74.

M. B. Schaefer, "Some Aspects of the Dynamics of Populations Important to the Management of the Commercial Marine Fisheries," *Inter-Amer. Tropical Tuna Comm., Bull.*, 1954, *1*, *2*, 25-56.

A. Scott, "The Fishery: The Objectives of Sole Ownership," *J. Polit. Econ.*, Apr. 1955, *63*, 116–24.

V. L. Smith, "The Economics of Production from Natural Resources," *Amer. Econ. Rev.*, June 1968, *58*, 409–31.

L. Van Meir, "An Economic Analysis of Policy Alternatives for Managing the Georges Bank Haddock Fishery," unpublished doctoral dissertation, Univ. Kan. 1969.

Economics of Production from Natural Resources: Reply

By VERNON L. SMITH*

One always writes a paper on the assumption that a certain background knowledge of principles is shared by the reader. When this is not the case, communication can become very protracted, and when there is the kind of imperviousness to understanding that is revealed in the comment by Richard Fullenbaum, Ernest Carlson, and Frederick Bell (FCB), communication may be impossible. But I will do what I can in this reply to explicate those aspects of my original paper, with which FCB are having difficulties:

I

When I write a cost function $C=\phi(x, X, K)+\hat{\pi}$, I take it for granted that the reader is familiar with the textbook theory of least-cost production under certainty, and will know without instruction that output, x, is variable only by virtue of input variability (FCB note that I nowhere mentioned input variability). In the context of commercial fishing, it is therefore implicit in $\phi(x, X, K)$ that the firm has a production function $x=G(z; X, K)$ where z is a vector of inputs, and the technological externality parameters X (fish stock) and K (number of firms, if there is harvest "crowding") are postulated beyond the control of the individual firm when the resource is unappropriated. As a least-cost producer the firm chooses z so as to minimize $C=wz$ (where w is a vector of input prices), subject to the constraint G, given (X, K). Under the usual regularity conditions, this process defines a set of theoretical input demand functions and, therefore, a lowest possible cost function $\phi(x; X, K, w)$. In general, the resulting cost is variable with output, and the parameters (X, K, w) hypothesized to be outside the firm's control.

I chose to simplify the presentation of the model by assuming that the K firms were identical. This need not mean, as believed by FCB, that the firm cannot change its size, output, and structure of operations in choosing among different input combinations (this technological variability is what is specified by G), but merely that all firms are identical and make the same changes. An example of a "firm" (which is a unit of organization) might be a vessel, complete with nets or traps, or it might be a mother ship with satellite fishing boats. These, and other inputs, are in general assumed to be variable, if not in the short run, then in the long run. The size and composition of vessels, gear, labor applied, fuel consumed, frequency with which traps are run, or days and lengths of days spent fishing are all in general, assumed not to be invariable and unresponsive to input prices, the biomass of fish and the number of fishing units (where crowding or gear congestion is important), but the degenerate case of no variability is not excluded. In the case of crowding, to be sure the size and structure of fishing units may be important as well as the number so that K may not capture all that is essential in crowding.[1] But, as I took pains to note, neither may X capture all that is important about the stock of fish, whose size (or age)

* University of Massachusetts. I am grateful to the National Science Foundation for research support.

[1] The number of firms, K, would also serve as a measure of the stock of capital only in the case of the example used in my original article, i.e., where a "firm" corresponds to a fixed "vessel." It is obvious that, except at a stationary equilibrium point, K would not also serve as an index of capital in cases where the development of the industry through time involved significant redesign of gear and vessels. In fact in an industry subject to wide changes in the stock of fish, one would expect flexibility in the design of vessels to command a premium so that the basic capital plant could accomodate wide variation in operating conditions (and even in the species harvested) by adjustments in operating inputs. My model would certainly require reinterpretation, and, I hope, would encourage reformulation, to deal with particular considerations that might vary among fisheries. It is neither constructive nor facilitative of research for FCB to propose changes which reduce rather than expand the flexibility of the original model, which was itself an extension of, not a replacement for, the classical literature. Neither that literature, nor my papers, are so rigidly defendable as "truth."

distribution may be an important determinant of cost (as well as fish reproduction). I hope it is not self deception for me to think that I have made myself clear!

II

So much for the general treatment of cost, which is of course essential to a general theory of natural resource production since different resources, even different fisheries may exhibit very different cost behavior. In the case of fisheries, FCB propose that very restrictive assumptions be imposed on the cost function. However, the reader should be warned that they are wrong in thinking that my model and my analysis does not fully include the situation they want to represent.

The case treated by FCB is precisely my model with a cost function given by

$$(1)\qquad C(x, X, K) \equiv \begin{cases} \hat{\pi}, & x \leq g(X, K) \\ \infty, & x > g(X, K) \end{cases}$$

The theory of the firm says that output is expanded as long as price is not below marginal cost. Since FCB are compelled by their assumptions to believe that marginal cost is zero up to the output given by $g(X, K)$, and infinite for greater outputs (i.e., larger outputs are impossible) this means that the firm's profit maximizing output is $x=g(X, K)$ which is consistent with my model, taking account of the corner conditions of maximization.

III

FCB's assumptions about cost are thus not precluded by my model, but the reader may wonder, as do I, why their faith in such a severely restricted cost function is held with such conviction.[2] They would have us believe, for example, that no matter how large the stock of fish, the firm continues to choose the same input combinations and incur the same total cost: that the fisherman will have no incentive to change his gear (size or number), to change his intensity of use of that gear, fuel consumption, etc. Their imagination is too limited to admit the possibility that a doubling of the stock of fish would not only increase output because the net or trap hauls are larger (a "windfall"), but also because less time is spent finding or chasing the fish or hauling in half-empty gear, so that an extra run to the fishery may be accommodated. It would be surprising if output, fuel, labor, and gear costs were not affected by all these considerations, but this is really an empirical matter on which FCB offer only dogmatic assertions that output and total cost cannot be varied by the firm. I hope that fisheries policy does not follow farm policy where it was once thought that removal of land from cultivation would be adequate to reduce agricultural surpluses, as if the intensity of fertilization and other aspects of cultivation were invariable.

IV

As noted above, the question of which special cost function is most appropriate is an empirical question. For empirical work I would suggest[3] a marginal cost function of the form indicated by curve A in Figure 1, rather than B as proposed by FCB. The hypothesis would be that the firm faces an upper bound on its catch rate given by $\bar{x}=g(X, K)$. No matter how intensively the firm fishes, it could not exceed a certain catch rate determined by the biomass and number of firms. But it can approach $\bar{x}$ more or less closely by varying inputs and therefore cost with marginal cost becoming steeper the nearer x is to $\bar{x}$. For example, the function

$$C' = 1/(\bar{x} - x)^{\alpha},\ x < \bar{x}, \alpha \geq 1$$

has this property. Furthermore, for $x<\bar{x}$, $C' \rightarrow 0$ as $\alpha \rightarrow \infty$ so that marginal cost B is a limiting case of A. If empirically one obtained very large estimates of α from firm

[2] A simple industry model like that of FCB has already been presented in James Quirk and V. L. Smith, (pp. 29–30), but I trust no one will view it as the last word in fisheries theory, instead of an illustration using restricted assumptions.

[3] I hope no one will take me too literally when I use the word "suggest." I would not recommend that serious empirical work begin until the fishery harvest process is stochastically modeled, since it is obvious to anyone who has ever wet a line, that an important feature of commercial fishing is the uncertainty of the catch. The firm's inputs, including (X, K) are most accurately thought of as parameters in a probability distribution function defined on harvest quantities.

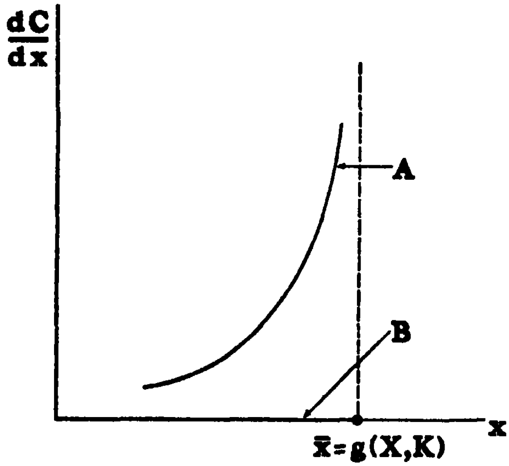


FIGURE 1

cost data, this would lend support to the FCB faith, otherwise not. What is crucial for empirical work is a cost hypothesis which does not exclude the possibility that cost will depend on x, X, K, and perhaps other variables, and which is rich enough to include special limiting cases.

V

I leave it to the reader to puzzle out the meaning of the FCB rhetoric in the four sentences immediately following equation (2).

VI

Equations (3)–(7) are described by FCB misleadingly as the "classical system," but, as the reader can verify, it is just a reproduction of my model using a special "classical" cost function (1). In fact they even employ the same phase diagram representation in Figures 1–4. The traditional literature, cited by FCB, does not employ these tools and does not state a dynamic system such as (3)–(7) modeling the bionomic interaction of industry with the resource. It is reassuring that except for the restriction to a particular form of the cost function, the rest of my original model is embraced warmly by FCB.

VII

FCB claim that my figures illustrating the set of points $F(X, K)=0$ corresponding to biomass equilibrium are "erroneous" in showing a positive slope, and that the "error is directly traceable to the misspecification of the cost function." Both statements are false.

As to the first statement, I showed on page 418 (also in the Appendix, page 430) that the sign $dK/dX|\dot{X}=0$ can be positive or negative depending on the signs of derivatives of the cost, revenue, *and* fish growth function, f. I further showed that if $C_{13}=0$, $C_{12}<0$, and $C_{11}>0$, then the sign is either (i) always negative or (ii) nonegative for all $X\leq\hat{X}$ (where $\hat{X}$ is a point such that the slope is zero). Surely it must be obvious from this that I was well aware that my model is fully consistent with $F(X, K)=0$ defining an inverse relationship for all (X, K). However, because the model showed the possibility of a positive slope, Figures 3–5 were drawn up so that "in each case it is assumed that (the sign) is positive below some value of X, and negative above that value." It would have been erroneous to do otherwise.

The second statement is wrong because their special assumptions about cost are neither necessary nor sufficient for the set $F(X, K)=0$ to have a negative slope. This can be proved by noting that their equations (3) and (4) define $\dot{X}=F(X, K)=f[X, Kg(X, K)]=0$. Hence,

$$\frac{dK}{dX}\bigg|_{\dot{X}=0} = -\frac{f_1+f_2Kg_1}{f_2(Kg_2+g)} \gtrless 0,$$

where $g_1>0$, $g_2<0$, $f_2<0$, $f_1\gtrless 0$.

FCB get the negative sign throughout by placing still more restrictions on their model (3)–(7). They use a homogeneous quadratic for $f(X)$, and set $g(X, K)\equiv rX$, with the result that $dK/dX|_{\dot{X}=0}$ is negative and constant. But this negative slope can be obtained with much less restrictive assumptions. Here is an example: Let $\dot{X}=aX-bX^2-Kx$, and $C=(\gamma x^2/X)+\hat{\pi}$. Then setting price equal to marginal cost gives $p=2\gamma x/X$, $\dot{X}=F(X, K)\equiv aX-bX^2-(p/2\gamma)KX$, and therefore when $\dot{X}=0$, we have $K=(2\gamma/p)(a-bX)$. The implication is clear. If the set $F(X, K)=0$ has a negative slope in the phase space, this implies *absolutely nothing* about any particular form for the cost function. What a negative slope does is to put a condition on the derivatives

of both C and f. If price is constant, the condition is (p. 418)

$$\left.\frac{dK}{dX}\right|_{\dot{X}=0} = \frac{KC_{12} + C_{11}f'}{xC_{11} - KC_{13}} < 0,$$

which should be the starting point for FCB if they want to explore the implications of an inverse relation between X and K.[4]

VIII

FCB are wrong in thinking that when $f'>0$, the policy implication is that the industry should be subsidized. The statement $\lambda = KC_2/f'$ does not hold (this is just elementary constrained maximum theory) *except* at a stationary state maximum point for the firm under appropriation, and $f'<0$ at a maximum. It follows that $\lambda>0$. When FCB try to interpret the tax as a subsidy "when" $f'>0$, they do so incorrectly on their own responsibility. I made no such interpretation, and I very much hope the fishing industry will never suffer from such misinterpretations.

Certainly under free entry we might have equilibrium with $f'>0$, but this is also not socially optimal and by no stretch of the imagination can one state that $\lambda<0$, calling for a subsidy. Starting from such an initial position, if the socially optimal equilibrium stock is obtained, say by placing a temporary moratorium on fishing, that stationary equilibrium could then be maintained by permitting free entry under the social charge, λ.

Now the question of how you get an industry to a maximum (the optimal time path), as distinct from how you keep it there after it has arrived, cannot, and was not intended, to be answered by a stationary state model. For this purpose one needs a model of maximization over time (see Quirk and Smith, and Oscar Burt and Ronald Cummings). In such a model the appropriate social charge will vary over time as the capital and fish stocks change, until a stationary equilibrium is obtained.

IX

It is not illuminating to say that "the classical theory can define a tax which depends upon the parameters but not the slope of the biological growth function." *Any* theory can do this if its assumptions are regarded as variables to be adjusted until that result is obtained! FCB get this result by treating the special case discussed above in Section VII.

X

The dynamic paths shown by FCB in their Figure 1 are not correct, i.e., such paths do not follow from their differential equation system. Refer to the derivation of the path conditions (A.3) and (A.4), and the phase diagrams in my paper.

[4] A fairly general set of sufficient conditions for a negative slope are the following: no crowding externalities ($C_{13}=0$); $f(0)=0$; and a cost function homogeneous of degree one. The cost function can then be written $C(x, X) = X\psi(x/X)$. Hence, $C_{11}=\psi''/X$, $C_{12}= -(x/X^2)\psi''$, and the above expression reduces to

$$\left.\frac{dK}{dX}\right|_{\dot{X}=0} = \left(\frac{1}{x}\right)\left[f' - \frac{f}{X}\right] < 0,$$

since $f(0)=0$, $f'(0)>0$, and $f''<0$ imply that $f'<f/X$, for $X>0$. That is, marginal growth in the fish stock is always below average growth. But $f(0)=0$ rules out the possibility of a critical species population below which it is not viable. As to cost homogeneity, I am skeptical because fishing units are confined to operate on a plane while many prey occupy a volume below. This strongly suggests a surface volume relationship underlying the physics of mass contact between predator gear and prey. Cost would then not be homogeneous since for given fish and vessel densities, the fish mass will vary roughly as the cube while fishing units will vary as the square of the radius (or other lineal measure) of the contact region.

Finally, bear in mind that all of this assumes price remains constant in the face of changes in the total fishery catch. If we drop this rather unrealistic assumpthen, then the above conditions are not sufficient to yield unambiguously a negative sign for dK/dX (see p. 430 of my original paper for the general case where price varies with the harvest).

REFERENCES

O. R. Burt and R. G. Cummings, "Production and Investment in Natural Resource Industries," *Amer. Econ. Rev.*, Sept. 1970, *60*, 576–90.

J. P. Quirk and V. L. Smith, "Dynamic Economic Models of Fishing," *in* A. Scott, ed., *Economics of Fisheries Management: A Symposium*, Vancouver 1970, pp. 3–32.

V. L. Smith, "The Economics of Production from Natural Resources," *Amer. Econ. Rev.*, June 1968, *58*, 409–31.

Employment and Rural Wages in Egypt: A Reinterpretation

By James A. Hanson*

Within the last two years a series of articles has been published dealing with the existence of surplus labor in the United Arab Republic. The most recent, an article in this *Review* by Bent Hansen, used the results of a rural employment survey conducted by the International Labor Organization and the United Arab Republic government (*Employment Problems in Rural Areas*, henceforth *EPRA*) to "settle the issue of surplus labor rather definitely against the surplus labor hypothesis" in that country (Hansen, p. 298). In the present author's view, neither Hansen's article nor the *EPRA* survey itself provide the evidence for such an unqualified rejection of the surplus labor-traditional wage hypothesis; nor do they suggest unqualified acceptance of a marginal productivity theory of distribution. Instead, it will be argued that the evidence is broadly consistent with more sophisticated hypotheses about rural wage determination. This latter view holds that with correct policy, labor might be drawn out of the agricultural sector for work on development projects or in regional industry, at relatively constant wages.

I. The Theory of Wage Determination in Peasant Economies with Abundant Labor

The main thrust of Hansen's article is contained in his statements "that farmers have a real choice between taking paid labor outside the farm or working on the farm. Under these circumstances there is little reason to believe that the value of the marginal product of labor, even on the smallest farms, would be smaller than current rural wages" (p. 302). This is particularly true for male adults who "seem by and large to be fully employed with long working hours during spring and summer and some underemployment from October to February" (p. 311).

Leaving to later sections the empirical question of whether paid, nonfarm employment is available, Hansen's theoretical argument that the marginal product of labor is positive and equals the wage is not strictly correct. Although the surplus labor theorists have never satisfactorily resolved the apparent contradiction between the existence of positive wages and zero marginal products,[1] it must also be admitted that the mere payment of wages does not imply competitive wages which are equal to the marginal product of labor. Particularly in the context of an underdeveloped, peasant, agricultural economy, a variety of noncompetitive distribution systems which involve wages are possible.

For example, assume that the rural economy largely consists of small farms, operated mainly for subsistence by families of homogeneous laborers. Next, assume that a laborer must be physically present to obtain his share of the implicit land rents earned by the family plot and that the offers of paid employment are some distance from the family plot. In this case, laborers would respond to the difference between the market wage and the average product of labor on the family farm, not the marginal product. Wages would, therefore, exceed the marginal product of labor on the family farm. The marginal product is not relevant because the economy has not developed institutions that permit the collection of property income without the need for physical presence. Such a model of a rural economy certainly has some applicability to Egypt, for the bulk of hired labor comes from small farms operated

* Brown University.

[1] For some attempts at resolving this problem, see John Fei and Alpha Chiang, and Harvey Liebenstein. Leibenstein's argument that hired labor must work harder and be paid more has been applied to Egypt by Donald Mead.

by large families (*EPRA*, II, Part II, Tables 37–60), and various observers have noted the long distances which workers need to travel to obtain employment (*EPRA*, I, pp. 54–55; Mead, pp. 96, 99).

Alternatively, one could assume that holding land and hiring labor confer social status, while working for wages reduces it. According to Mead and H. Habib Aryout, two commentators on rural Egypt, these views are common among Egyptian peasants. Under this assumption, workers could be obtained only by paying wages in excess of the marginal product of labor on subsistence farms; while employers would be willing to pay more than the marginal product of labor on their farms to gain the "consumption aspects" as well as productive benefits of labor. The social preference for self-employment would lead to a wage rate which exceeded the marginal product of labor on subsistence farms. Furthermore, the social preferences might manifest themselves in other ways: even small farms might hire extra labor for "consumption purposes," very small farms might consist of separate, unconsolidated plots in order to prevent the owner's neighbors from learning of the small size of his holdings and his need to work for pay, and peasants might travel considerable distances to seek employment, rather than face the degrading need to work for their own peers. All of these corollaries of the positive preference for self-employment are prevalent in Egypt.[2]

If we ignore the possibility or importance of "market imperfections," highlighted by the above arguments, Hansen's argument is correct, but it may be irrelevant. Although the peasants may actually equate on-farm marginal products with off-farm earnings, there may be few job offers during the slack season, October to February. Thus the average farmer would be forced to occupy himself with low productivity, nonagricultural pursuits during this time. Correspondingly, employment opportunities for women and children who engage in these activities in the busy season would be greatly reduced. This argument has been hinted at by Mead and by Hansen himself (pp. 304, 311). The remainder of this note is devoted to providing evidence which supports this view and, incidentally, casts doubt on Hansen's first point that wage labor is a viable alternative to on-farm work.

[2] See Hansen, (p. 303); *EPRA*, I, (pp. 54–56, 67–70); II, Part II, Tables 37–67; Mead (p. 97).

II. The Extent of Unemployment and its Seasonality

Hansen's argument that surplus labor is nonexistent and the labor force is fully employed is based on a special interview subsurvey taken during one week in January, at the depth of the seasonal trough (Hansen, pp. 304–06; *EPRA*, I, pp. 157–61). However, the quoted unemployment rate of 8 or 9 percent during this slack period is actually lower than the 12.5 percent average ***annual*** rate among adult males in agricultural enterprises which is cited in the complete survey (*EPRA*, I, p. 183). Since the methodology that was used to calculate these two conflicting figures leaves something to be desired,[3] we have turned to the basic data re-

[3] The figures cited by Hansen come from a special labor survey taken during one week in January when unemployment is highest (*EPRA*, I, pp. 157–61). According to their responses potential workers were divided into jobless (52 percent) and working (48 percent). Of the jobless, 81 percent were unemployed for reasons of school attendance, old age, physical handicaps, or military service. One percent were employers who were regarded as working, leaving 18 percent of whom 4 percent were workless because of lack of opportunity, 9 percent because of tradition, and 5 percent who did not reply. The division of these figures into an unemployment rate of 8 or 9 percent of the labor force is rather hazy at best. Nor does the rate include underemployment of those persons actually working during the reference week. These factors might easily yield figures of well over 10 percent unemployment, which correspond to those obtained from the full survey's hourly data described in the text.

As to the figure of 12.5 percent unemployment contained in *Report C* of *EPRA*, I, there is no accurate description of which subset of the full survey was used to obtain the figures for hours worked by provinces and regions, nor the methodology by which it was selected (*EPRA*, I, pp. 214–23; based on *EPRA*, II, Part II, Tables 1–18). Using the methodology of *Report C* to estimate unemployment (303 eight-hour days), no reasonable subset of labor input by strata taken from the full survey (*EPRA*, I, pp. 224–228; II, Part II Tables 37–65) gives unemployment rates that approach the 12.5 percent figure. Finally, when the regional figures

ported in *EPRA* and calculated our own annual unemployment rates.

According to the survey tables (*EPRA*, I, pp. 100–02; II, Part I, pp. 58–67), men[4] were completely unemployed an average of 70 days while men in agricultural households were idle 81 days. Assuming, following the survey, that workers rest on Fridays and holidays, there are 303 possible working days (*EPRA*, I, pp. 178–79) and the resulting unemployment rates are 5.6 and 6.3 percent, respectively. Turning to the hourly figures, on 28 of his working days the average man worked less than 6 hours, while the average man in an agricultural household worked less than 6 hours on 24 days[5] (*EPRA*, I, p. 101; I, Part I, pp. 62–63). However, on other days they were often employed for more than 8 hours, and depending on which subset of the survey is chosen, the average workday ranges from 8.5 to 8 hours. Therefore, if we again follow the survey and assume that an 8-hour day is the norm,[6] the rates of unemployment based on hourly data are about the same as, or sometimes much less than those based on days worked. As Hansen points out, these results are only slightly lower than the unemployment percentages obtained through the interview survey during the slack season and thus support his hypothesis of full employment.

However, in any estimate of the stock of unemployed labor the norm must be chosen with care and in this case the norms seem much too low. First, it is likely that many more than 303 days are available. Another special subsurvey, taken in April, shows no difference between the number of hours worked on Fridays and other days. Moreover, the major religious holiday, Courban Bairam, was observed with only 1 day of rest, rather than the 4 assumed in the survey. In the words of the survey: "In practice, they (the farmers) work if there is work available and they rest if there is a lack of work."[7] Correcting only for the understatement resulting from the holidays of Courban Bairam and Ramaden Bairam would add roughly 2 percentage points to the unemployment rate.

The survey's norm of 8 hours worked in each of these days is also too low and seems to be based on Western industrial practices rather than non-Western peasant agriculture.[8] Its inapplicability becomes particularly obvious when we turn to a discussion of the seasonality and regional factors in unemployment. Although there is a variation depending on the type of crop, broadly speaking, there is a busy season with intermittent lulls from March through September with a slack season after the last crop, cot-

that were used to calculate the 12.5 percent figure are divided between Upper Egypt and the Delta, there is no significant difference in employment rates. All other evidence contradicts this view. Thus it seems best to use the full labor survey rather than this undocumented subset.

[4] In this case men refers to adult males who are permanent members of a household. Workers are regarded as permanent if they spend over thirty days working for the same farm. Although no temporary workers per se are included in survey, there are two strata of landless farmers. Since their unemployment rates are higher, to the extent that the survey includes too few of this group the unemployment figures will be too low. Figures for men are used here because the availability of children and the division between household and farm work are uncertain. Moreover, in a traditional, nonmatriarchal society unemployment among men should indicate even greater unemployment among women and children. However, these classes are not necessarily reservoirs of labor for, as income rises and educational opportunities increase, these groups may withdraw from the labor force. Counteracting this tendency is the decline in tradition mentioned by Hansen and the shift from production of household goods, or subsistence handicrafts to specialization and purchase in the market with a decline in the time necessary to maintain the house. The causes and effects of this shift have been described recently by Stephen Hymer and Stephen Resnik.

[5] Figures for average number of days worked and average annual hours worked refer to slightly different groups of laborers.

[6] The survey claims that it used a 7-hour day (*EPRA*, I, p. 179) but the actual figures for weekly and annual available labor are divisible by multiples of 8 and not 7 hours (*EPRA*, I, pp. 212–28).

[7] See *EPRA*, I, (pp. 107–109, 182). If we assume, following the survey, that there are only 3 working days of 8 hours in the survey period containing Courban Bairam and 4 working days of 8 hours in the period containing Ramadan Bairam, the overemployment is 213 and 129 percent for all workers. The corresponding figures for agricultural workers are 187 and 116 percent (*EPRA*, I, pp. 182, 214, 217).

[8] On the problems of applying Western norms see Gunnar Myrdal's *Asian Drama*, ch. 21, 22, and Appendix 6.

ton, is harvested. During this peak season, the average number of hours worked increases significantly, from 6.4 to 8.8 per day.[9] Assuming that enough calories are available to make it physically possible to work at the peak rate throughout the year, this fact implies that a substantial number of hours are available in the present slack period for work on a second crop, infrastructure, or in cottage industry; hours which could be worked without interfering with the present production of food or cash crops. In fact, on this basis, and correcting for the errors in the holiday weeks described above, annual unemployment among permanent male workers in agricultural enterprises amounts to 10 or 11 percent, almost double the uncorrected survey figure of 5 to 6 percent.

The large seasonal variation in employment also casts doubt on Hansen's contention that workers have a real choice between taking paid labor and working on the farm. In fact, although paid employment outside the household enterprise represents 20 to 25 percent of the men's work time, 65 percent of this work occurs during the busy season (*EPRA*, I, p. 69; II, Part II, p. 68, Tables 18.a, and 37–60).

Of course, one could argue that workers are idle in the off-season by choice rather than through the lack of employment offers. In that case, there is no surplus labor, for potential employers would have to overcome the workers' preferences for leisure through higher wages and few additional workers could be obtained at the going wage rate. To settle this point a knowledge of the response to wages, which is discussed in the next section, and, ultimately, preferences for income and leisure is necessary. However, there is some indirect evidence that measured unemployment is due to lack of demand for labor rather than supply factors.

According to the survey a great disparity exists between employment rates in Upper Egypt, which has only one crop a year in most areas, and the Delta, where irrigation and two crops are more prevalent (Hansen, p. 303; *EPRA*, II, Part II, Tables 10A, 12A, B). Assuming regional attitudes toward work are similar, this suggests that much of the unemployment is technological (and regional).[10] Without the irrigation which makes a second crop possible, there is little to do in Upper Egypt during the off-season. Throughout the country a large percentage of working time during that period is spent on nonagricultural and nonfarm jobs such as animal watching, marketing, and production of handicrafts (*ERPA*, I, p. 151 and Graph 20) that have an obviously low productivity. It seems unlikely that much time would be spent in this way if, as Hansen contends, wage employment existed. If more remunerative or productive jobs were made available in the off-season, this low level work might be shifted to women or children, reducing their unemployment. As Hymer and Resnik have pointed out elsewhere, this shift to more productive work would be even more likely and wages would rise even less, if the output of the new jobs was a good sub-

[9] See *EPRA*, I, (pp. 110–54). A statistical test of this difference is easily made. Assume that the total weekly hours worked by those sampled in each of the strata (*EPRA*, II, Part II, Tables 37–65; I, pp. 212–28) are a random variable, influenced by changes in weather, etc. (Owing to the problem of Fridays discussed in the text, it was felt that total weekly hours was the relevant variable. However, to maintain consistency with the previous paragraphs, figures in the text were calculated by excluding holiday weeks and assuming no work was done on Fridays.) Under this assumption the difference between the average hours worked per 8-day period in the busy and slack season is statistically significant at the 5 percent level, for each of the 8 strata of permanent male agricultural labor. The strata, listed in Hansen's article, are the result of cross classification by farm area and family size.

[10] Mead's calculations support this result. As Hansen points out (pp. 303–04, 311), there is an even greater regional disparity in unemployment rates among women and children, so technological unemployment is underestimated by this calculation. Of course, these disparities may be due to regional differences in traditions and attitudes toward work and leisure, but traditions change, and, if the change occurs through greater offers of work at the going wage we have a surplus labor situation. One discordant note does arise over the use of this data which is from the full survey but based on labor inputs. The regional employment data contained in the full survey and described in fn. 3 do not show a similar variation between regions. As discussed in that footnote the regional data is judged to be unreliable, but if the reader decides it is preferable, he also must accept the resulting unemployment rate, 12.5 to 15 percent.

stitute for the production of subsistence handicrafts.

Another test of the hypothesis that unemployment is due to choice rather than technology is a comparison of hours worked per permanent worker on different size farms; if employment were readily available then there should be no difference in the time worked per permanent worker. However, the results are dominated by farm size; in all cases save one, the larger the farm, given the family size, the more total hours worked by the average permanent male worker. Also, seasonal fluctuations in employment are much larger on the smaller farms, mainly as a result of the seasonality of outside work (*EPRA*, I, pp. 118–120). These results seem to indicate that additional work on small farms runs into diminishing returns, and contrary to Hansen, outside work is not always available.[11]

III. Rural Wages

As further support for the hypothesis of marginal productivity distribution, Hansen presents wage data from the survey (*EPRA*, III). Both the method of compiling the data—asking the village sheik for prevailing wages—and the procedure of excluding villages in which a substantial number of observations were missing would seem to bias the result toward the traditional wage theory. However, according to Hansen, the distribution of the wage data is not truncated, or squeezed up against a traditional subsistence minimum, but fairly normal, disproving the traditional subsistence wage theory once again.

Of course, as shown earlier, disproving the traditional subsistence wage theory is not the same as proving that distribution is by marginal products. Moreover, Hansen's argument against the traditional wage theory is faulty on at least 3 counts. First, out of the original sample of 48 villages, 17 were eliminated as having missing information, particularly during the slack season (Hansen, p. 306; *EPRA*, III). This fact casts still further doubt on Hansen's contention that nonfarm wage labor is a viable alternative to farm work throughout the year. Instead it supports the view, expressed earlier, that there is large seasonal and, since most of the missing villages are from Upper Egypt, regional technological unemployment. Second, only the most naive proponent of the subsistence wage theory would argue that workers react to money wages; the argument is usually phrased in terms of the money cost of a subsistence bundle of goods. However, the wage data presented by Hansen are undeflated by regional prices. The substantial variation in average money wage between govornerates, particularly between Fayoum, an oasis in Upper Egypt which is separated from the Nile Valley by 10 miles, and the rest, is certainly evidence of some imperfect connections between factor and/or goods markets. Third, if we decide to use the data in spite of this obvious deficiency, but omit the 6 villages in Fayoum, which are the only observations from Upper Egypt, the distribution appears truncated rather than normal. Statistically speaking, neither the hypothesis that the data are normally distributed, nor the hypothesis that they are distributed as a truncated half of the normal distribution can be rejected at the 5 percent level.[12] Thus this evidence cannot be used to

[11] An analysis of variance test of average hours worked per man, based on a table classifying farms by area and family size, shows the differences in labor employed in different size farms are significant at the 10 percent level. The data are the calculated average annual hours worked by permanent male employees in each strata and were taken from the survey (*EPRA*, II, Part II, Tables 37–65 and *EPRA*, I, pp. 212–28). These results also show up in Hansen's labor input data (Table 2, p. 301) and were mentioned in the survey (*EPRA*, I, pp. 118–20). Again the result might be due to a difference in tastes, since some farmers may have been able to purchase larger farms because they were harder workers and accumulated more wealth.

[12] A Kolmogorov-Smirnov test of the goodness of fit to a hypothesized distribution, as described in B. Lindgren and G. McElrath, was performed on the observations from the remaining 25 Delta villages (Hansen, Table 5, p. 297). Under the null hypothesis that the sample was drawn from a normal distribution with mean and standard deviation of 18.79 and 2.71 piasters per day, respectively, the largest difference between the cumulative frequency of the sample and the hypothesized normal distribution is .20. This is a smaller difference than we would get under the null hypothesis 95 percent of the time and so we cannot reject the null hypothesis. For the truncated distribution the observed

reject either the marginal productivity theory or the traditional wage theory of distribution.

Even more puzzling is Hansen's discussion of variations in wages. He states that there is substantial seasonal fluctuations in daily wages, since the highest daily wage for men is 150 percent of the lowest, and he uses this variance to support his claim of great flexibility in wage determination. However, no other figures are presented on the variation. Moreover, part of this variation could obviously be accounted for by the inflationary trend in the case of the subsistence market basket and part by the variation in the average number of hours worked per day.[13] In fact, Hansen's regression equation of average daily wages on hours and a time trend is

$$W_t = 5.630 + \underset{(.023)}{.219H} + \underset{(.008)}{.073t} \qquad R = .88$$

where

W_t = average daily wages of men at time t
H_t = average hours of work per week (all men)
t = time t = (0, 1, . . . , 52)

and all coefficients are highly significant. Hansen interprets the regression as supporting the hypothesis of demand and supply determined wages by identifying average hours of work per week as a proxy for days worked and assuming that demand shifts while the supply curve has a positive slope and is fixed.

It should be clear however, that there is another interpretation of the regression equation which strongly supports the traditional wage-surplus labor view. The alternative hypothesis, consistent with the labor surplus theory of wage determination and supported by the regression, is a horizontal or nearly horizontal supply curve of labor (measured in hours) at some constant real hourly wage.[14] Neglecting the constant, for which no standard error has been provided, and assuming the time trend represents only inflation, the regression coefficient for weekly hours, divided by a constant equal to the available man-days per week, is equal to the constant, average, real, hourly wage.[15] Under this hypothesis average daily wages would then be determined by shifts in the demand for labor hours against a horizontal supply curve of labor, also measured in hours.

The hypothesis of a horizontal or nearly horizontal supply curve is also supported by some alternative calculations. The annual difference between hourly wages in the two overlapping weeks (June 21 and 28, 1964 and 1965) in the wage survey is 37.5 percent and 37.3 percent, respectively (*EPRA*, III, p.

frequency table of daily wages was inverted between the second and third classes (containing 2 and 7 observations), and the resulting set of new and sample observations were summed by classes. This procedure results in a normal distribution which is symmetric about a mean of 15.95 piasters per day, with standard deviation 3.92. Under the null hypothesis that the sample was drawn from the right tail of this distribution, we find that the largest difference between the cumulative frequency of the sample and that expected under the null hypothesis is also .20. Again we cannot reject the null hypothesis at the 5 percent level of significance. In other words, the available data does not permit the investigator to distinguish between the two hypotheses.

[13] Hansen assumes laborers are hired on a daily basis and there is little seasonal fluctuations in hours worked. He admits this point is crucial to his interpretation of the regression (p. 309). However, as shown earlier, the seasonal fluctuation in daily hours worked is significant.

[14] There is no reason to think that Hansen's daily wages rather than hourly wages are the correct price variable. One could easily imagine a traditional hourly wage. If leisure has any value and workers can vary their hours, hourly wages are the correct variable. As R. Albert Berry and Ronald Soligo have shown, except in the case of leisure satiation this implies a rising supply curve of labor, although the exact elasticity is not discussed. Our argument in this paper implies that the elasticity of labor supply is very high. Another possibility which would result in a horizontal labor supply curve is constant labor productivity in the production of subsistence handicrafts, which, according to Hymer and Resnik, are the real alternative to field work.

[15] If the constant is statistically significant it might simply be a statistical reflection of an accounting or traditional tendency to hire laborers for some fixed number of hours per day. This proposition could be tested either by regressing daily wages against a constant and the average number of hours worked in excess of the fixed number, and testing for the equality of hourly wages for regular time and overtime, or most simply by regressing hourly wages against a constant, hours, and the time trend. A rough version of the second test was performed in the text.

37). Ignoring growth in the labor force, and making Hansen's assumptions about the equality of time worked during the same periods in the year, this figure represents the inflationary trend.[16] A rough series of real hourly wages can then be obtained by deflating each of the 54 weekly average hourly wage figures in the 31 village sample by the inflationary trend. Although the real hourly wage is 10 to 15 percent higher in the busy season—March to September—if we treat the weekly observation as a random variable, this difference is not significant at the 5 percent level. Thus we accept the hypothesis of a constant, traditional hourly wage. On the other hand, as discussed earlier, there is a statistically significant seasonal variation of 25 to 30 percent in weekly hours worked by males in each agricultural strata. This supports the view of a shifting demand curve for labor intersecting a relatively horizontal supply curve.

IV. Summary and Policy Implications of the Alternative Hypothesis

This note has argued that the evidence presented in Hansen's article and the *EPRA* survey on which it was based is broadly consistent with either a marginal productivity theory of distribution, or a traditional subsistence wage theory of distribution. However, the evidence is most consistent with a more sophisticated hypothesis of wage determination incorporating both substantial seasonal variation in employment and handicraft production.

What are the policy implications of accepting the more sophisticated hypothesis rather than the marginal productivity hypothesis of wage determination For example, what do the different hypotheses imply about the impact of the Aswan Dam? Should the Egyptian development effort be biased towards large scale manufacturing, smaller scale, regional dispersed manufacturing, or agricultural improvements?

To answer the first question, the more sophisticated view, encompassing the seasonality of productive labor and the variations in employment by regions and size of holdings, seems to indicate the possibility of really significant increases in food output through the greater use of irrigation, which would make two-crop rotation possible in Upper Egypt, the development of crops which mature in shorter seasons, and the use of fertilizers to prevent declining soil fertility. At present, five weeks are lost in January owing to the lack of water. The Aswan Dam will make this loss unnecessary (*EPRA*, I, pp. 117–18) and if crops can be modified to mature in a shorter time, three crops may be possible, provided increased amounts of fertilizer are used to prevent declines in soil fertility. If workers were presently fully employed throughout the year in productive labor, as implied by Hansen's marginal productivity theory, such output increases would be much less.

To answer the second question, there seems to be some possibility of building infrastructure, transportation networks and then developing regional industry using off-season labor or the present landless peasants; this would be particularly true if the output of these projects yielded substitutes for the present uses of off-season labor, e.g., more durable irrigation works and farm machinery. In this way, output and capital would be increased without a decline in food production. Based on the combined results of the wage and employment survey described in Sections II and III, rather than a marginal productivity theory of income distribution, wages would not have to be increased much above prevailing levels to obtain the necessary labor for these projects.

[16] Hansen's figure, calculated from the regression, is about 30 percent.

REFERENCES

R. A. Berry and R. Soligo, "Rural Urban Migration, Agricultural Output, and the Supply Price of Labor in a Labor Surplus Economy," *Oxford Econ. Pap.*, July 1968, *20*, 230–49.

J. Fei and A. Chiang, "Maximum Speed Development through Austerity," in I. Adelman and E. Thorbecke, eds., *The Theory and Design of Development*, Baltimore 1966.

H. Habib Aryout, *The Egyptian Peasant*, Boston 1968.
B. Hansen, "Employment and Wages in Rural Egypt," *Amer. Econ. Rev.*, June 1969, *59*, 298–313.
S. Hymer and S. Resnik, "A Model of an Agrarian Economy with Non-Agricultural Activity," *Amer. Econ. Rev.*, Sept. 1969, *59*, 493–506.
C. Kao, K. Aschel, and C. Eichner, "Disguised Unemployment in Agriculture: A Survey," in C. Eichner and L. Witt, eds., *Agriculture in Economic Development*, New York 1969.
H. Leibenstein, *Economic Backwardness and Economic Growth*, New York 1963.
B. Lindgren and G. McElrath, *Introduction to Probability and Statistics*, New York 1959.
D. Mead, *Growth and Structural Change in the Egyptian Economy*, Homewood 1967, pp. 90–98.
G. Myrdal, *Asian Drama*, New York 1968.
Institute of National Planning, *Employment Problems in Rural Areas*, U.A.R.
(I) *Report C on Utilization of Rural Manpower and Measurement of its Underemployment*, prepared by M. Mohiey El-Din Nazart, S. Shabana, S. Rofael, U. Planck, and Q. Qayoum, Cairo, Aug. 1966.
(II) *Statistical Tables, The Labour Record Sample Survey*, Part I, Cairo, Dec. 1965; Part II, Cairo, Sept. 1966.
(III) *Report D on Wages, Income, and Consumption in Rural Areas*, prepared by B. Hansen, A. Sedki, and Y. Moustafa, Cairo, Dec. 1965.

Employment and Rural Wages in Egypt: Reply

By BENT HANSEN*

James Hanson has reinterpreted the data on employment and rural wages in Egypt which were collected by the rural employment survey (*Employment Problems in Rural Areas* (*EPRA*)) of the International Labor Organization and the Institute of National Planning, Cairo. Moreover, he has presented some qualifications to my interpretations which are justified, but his basic alternative interpretation of the seasonal fluctuations of daily wages does not hold water. This alternative interpretation was, in fact, carefully examined and rather effectively rejected by the *EPRA* report itself on the basis of information which Hanson, unfortunately, has overlooked.

Before I enter upon details, I want to make it clear—Hanson's exposition of my views gives the opposite impression—that there is no disagreement about the existence of some *seasonal* unemployment and underemployment, and the possibility that there may be slack periods during the year with temporary zero marginal productivity of labor in agriculture. I summarized my article (1969, p. 311) by saying that "*Male adults* seem, by and large, to be fully employed with long working hours during spring and summer, and with some underemployment from October to February"; that for the year as a whole "*Female adults* . . . are slightly underemployed . . . "; . . . "children between six and fifteen work, on the average, slightly more than women outside the households." I found it rather meaningless, although, to apply the notion of underemployment to children between six and fifteen (if anything, I found them overemployed). Thus I did not argue, as Hanson maintains, that surplus labor is nonexistent and the labor force is fully employed. When I said (p. 302), that "Farmers (i.e., men), by and large, are fully occupied, taking into account the work outside their own farms," this was a statement which referred to total hours worked per year without considering the distribution during the year. Nowhere in my article did I maintain that work outside the farm is a viable alternative to farm work throughout the year i.e., at any time of the year. Hanson's efforts to prove that men are to some extent unemployed at certain times of the year and that outside work is not always available only repeat what I thought was obvious from my article.

As I see it, the major issue in the rural surplus labor problem is: does there exist a surplus of labor in the sense that marginal productivity of labor at all times of the year is zero (or at least, very low) so that labor could be withdrawn permanently from agriculture without detrimental effects on output; or, is the surplus of a purely seasonal nature and if so, how large is the seasonal surplus and how long a time of the year does it exist. It is the first case which traditionally has been considered in development theory and which makes demand-supply theory of wages break down so that if wages actually are paid, some sort of "institutional" wage explanation is called for.[1] The existence of seasonal slacks and seasonal underemployment gives rise to entirely different development problems and is fully compatible with demand-supply theory of wages even in a fully competitive labor market and even if marginal productivity occasionally during the year should become zero. I shall discuss this more in detail below. The *EPRA* shows unequivocally that surplus labor in Egypt is a seasonal problem and of much less importance than has hitherto been assumed. Hanson has made no attempt to deny that.

* Professor, University of California, Berkeley.

[1] The problem of wage determination exists also in models which assume that rural employment and production is determined at the margin where utility of income and disutility of labor are equal so that marginal productivity is positive, but a gap is assumed to exist between marginal productivity in family enterprises and rural market wages. (Amartya Sen, p. 31 ff.)

I. Market Imperfections

Hanson points out that the fact, disclosed by the *EPRA*, that small farmers to a large extent find work outside their own farm does not really imply, as I contended, that "there is little reason to believe that the value of the marginal product of labor even on the smallest farms, would be smaller than current rural wages" (1969, p. 302). Preferences for self-employment may make farm families work on the farm to a point where marginal productivity is lower than current market wages. I agree in that. What makes it difficult to believe that marginal productivity generally should be lower than market wages even on small farms, is rather the fact that even small farms hire labor. We probably agree that when labor is actually hired it is hard to believe that marginal productivity should be lower than the market wage rate. Since even the smallest farms to some extent depend on hired labor (1969, p. 302, Table 3) there must be at least some time of the year when marginal productivity and market wages are in line with each other. Outside the periods when labor is hired, the preference for self-employment may, however, lead to a lower marginal product of labor than the market wage rate. The problem is then how important such preferences are. I still believe that the fact that up to 33 percent of men's work (namely on farms of ½ to 2 feddan with more than 3 working family members) is outside their own farm, indicates that such preferences cannot be too difficult to overcome.

Another "model" (the Lewis model) mentioned by Hanson, which leads to comparison between market wage and average rather than marginal product, overlooks the fact that in an environment like rural Egypt, a decision to take temporary work outside the household, is a family decision (or, rather, a decision by the family head) and not the decision of the individual family member; for that reason it is likely that the decision is taken on the basis of the effect on *total* family income, and in that case it is, of course, the comparison between market wage and marginal product that matters.

Let it finally be pointed out that even though preference for self-employment may imply that farms that do not rely on hired labor may have a lower marginal productivity of labor than current rural market wage rate, this does not in any way upset traditional demand-supply theory (Sen, p. 33); there will still be equality between wages and the marginal product of hired labor, or, to put it another way, market wages will be equal to marginal product plus possible "marginal self-satisfaction."

II. Estimated Underemployment

With respect to estimated underemployment, and thus surplus labor, two methods should be clearly distinguished: actually observed employment may be compared with either some social norm for employment, or the "intended" supply at the going return to labor or market wages. Only the first method was applied in the case of the *EPRA*, and conceptually the second method may not even be well-defined in a market with demand-supply determined wages. (See below.)

The norm adopted by the *EPRA* was 8 hours work per day with one weekly day of rest and a few religious feast days off. It implies 303 eight-hour days per year and average annual underemployment for adult males in agriculture of about 6 percent. Hanson considers this norm too low and calculates his own underemployment percent on the basis of 306 working days (if I have understood him correctly) with a daily number of hours equal to the actual average at the seasonal peak; on this basis the average underemployment percent for the year becomes about 10 to 11. Hanson accuses the norm of the *EPRA* of being based on Western industrial practices rather than non-Western peasant agriculture. Social norms are, of course, always arbitrary. The norm chosen by the *EPRA* was largely based on what was considered to be the social norm of the present Egyptian regime, which in 1962 introduced a 42 hour week with some paid vacation in the "organized" sector and generally respects all religious feasts. From a planning point of view it is natural to consider the actual norms of the government of

the country in question. The weekly day of rest (modern Western industrial practice is to have 2 such days) is, incidentally, an ancient rule, laid down in more than one major religion, thus presumably expressing the social norms of primitive non-Western peasant or tribal societies. In addition, it was taken into consideration that the average number of hours actually worked per day by hired, actually employed rural laborers turned out to be rather constant during the year June 22, 1964, to June 21, 1965 at a level of 8.1 hours. Hanson's norm seems to be based on the assumption of a given constant institutional wage rate and a horizontal supply curve (at least) to the point of actually observed peak employment. We shall see below that there is no justification for Hanson's assumption of a constant wage rate; the available evidence proves the opposite. In addition I would like to stress that even on Hanson's norm, underemployment in Egyptian agriculture is a purely seasonal phenomenon and much more limited than has usually been assumed. (The first Egyptian 5 year plan worked with the assumption that almost 25 percent of the rural labor force could be permanently removed from agriculture.)

In a further effort to bring up the average underemployment percent, Mr. Hanson reveals a peculiar, though by no means uncommon, prejudice against certain types of rural activities. I quote: "Throughout the country a large percentage of working time during [the off-season] is spent on non-agricultural and nonfarm jobs such as animal watching, marketing, and production of handicrafts . . . that have an obviously low productivity. It seems unlikely that much time could be spent in this way if, as Hansen contends, wage employment existed. If more remunerative or productive jobs were made available in the off-season, this low level work might be shifted to women or children, reducing their unemployment" (p. 495). Quite apart from the fact that "animal watching" (which is not a special Egyptian form of rodeo with which the rural population entertains itself during slack seasons, but simply animal husbandry which goes on all the year round) already to a large extent is the work of women and children, I do not understand when this kind of work has become "nonagricultural." Nor do I understand on what grounds animal husbandry, marketing, and domestic handicrafts (which in Egypt mostly means repair work and house building—rural families in Egypt usually build their mudbrick dwellings themselves) are deemed "low level work" with an "obviously low productivity," which ought to be left to women and children unless, of course, Hanson has in mind the fact that off-season marginal productivity of field work is relatively low and that any off-season activity should be expected to have low marginal productivity. Hanson does not seem to have understood that one of the important results of the *EPRA* is to demonstrate that in concentrating on field work only, earlier studies of rural employment missed a number of time consuming activities without which a traditional Egyptian farm simply cannot be run. Animals (some of them even necessary for the field work) have to be taken care of, products have to be sold and inputs purchased, and both transported, etc. And it is good to have a house to live in, even in the splendid Egyptian climate. If alternative high-productivity off-seasonal activities were created (requiring investments, presumably) farmers would, of course, shift away from both nonagricultural *and* agricultural work. But what has this to do with the problem at hand?

III. Rural Wages and Their Seasonal Behavior

Considering the great variability of daily wages, sex-age differentials between daily wages, etc., it is impossible in the collected data on daily wages to find any kind of constancy which could be interpreted as a non-market determined institutional wage level. Hanson believes, however, and this is his basic reinterpretation of the data, that seasonal variability of the daily wage rates is the outcome of a similar variability of the number of hours worked daily by employed

laborers at a relatively constant hourly wage rate. And in this constant hourly wage rate, Hanson believes to have found the missing institutional wage rate.

This interpretation should not be ruled out a priori, of course. The *EPRA* were, however, fully aware of this possibility. Information was therefore collected not only about daily wages, but also about the number of hours to be worked at the reported daily wage. Parallel with the weekly series for daily wages, the *EPRA* thus had at its disposal a series for hours to be worked and this series turned out to be very constant with at most a very slight seasonal movement. I mentioned this in a footnote to my article (1969, p. 309) which Hanson has noticed, but misunderstood. He seems to believe that I arbitrarily assumed that number of hours for employed laborers is constant, and that he just as well can make the opposite assumption. My assumption was, however, based on firm evidence which Hanson has overlooked completely. He did not trouble himself by looking up the reference given in the footnote just mentioned. The raw weekly averages for both daily wages and daily hours for men can be found in the *EPRA Report D*, p. 39: see also graph I:4a, and figures for calculated hourly wages were published in the same volume, p. 37. The text of *Report D* discussed the problem in some detail:

> Thirdly, an important point: as an expression for the demand for labor we have used the number of hours worked on average per eight day period per member of the labor force, while as an expression for the price of labor we have preferred to use the *daily* wages, partly because the information of daily wages may be more reliable than that of hourly wages. The parallel movements of wages and hours worked per eight day period might therefore simply express the same movements for hours worked for *actually* employed people at a constant wage per hour. Now, our earlier findings actually showed a maximal seasonal spread of wages per hour of the same size as that for daily wages, and a much smaller maximal spread for hours for actually employed laborers (see Tables I:1 and I:7). It seems unlikely, therefore, that the common fluctuations in hours worked for actually employed persons [read:laborers] should explain the correlation between daily wages and average hours worked per eight day period for all members of labor force. But hours per day for actually employed laborers do show some fluctuations, and in order to make sure that the seasonality of hours worked for actually employed [laborers] does not account for the correlation between daily wages and hours worked per eight day period for all members of the labor force, we have for ***men*** calculated the average number of hours worked for actually employed [laborers], and through simple division of average daily wages and average hours actually worked the average hourly wages were obtained week by week. The upward trend in hourly wages was eliminated in the same way as for daily wages, and the "true" seasonal fluctuations of hourly wages were then obtained as the difference between the trend and the actual weekly averages. See graph I:4a. The resulting adjusted hourly wages have been depicted in graph I:5 and it turns out that the index of (adjusted) hourly wages follows the index of (adjusted) daily wages very closely. This shows that movements in the average hours actually worked per day are mainly erratic; looking at the curve for the weekly averages of hours per day in graph I:4a, one also gets the impression that there is at most a very weak seasonality in the movements of hours per day. [pp. 29–30]

To permit the reader to judge for himself, I have amended Figure 1 of my 1969 article so that it now shows not only daily wages, and average working hours (per 8 day periods) for all men in labor force, but also the average working hours (per 7 days, to make the levels easily comparable) for actually employed laborers. The difference between the two employment curves is striking. Hanson's mistake was thus to overlook that information about hours for

actually employed laborers did exist, and to identify hours worked by actually employed laborers with the average of hours worked by all men in the labor force whether employed or unemployed, whether self-employed or employees. The reader will also understand now that I was fully justified in using the latter series as an index of labor demand to explain the fluctuations in daily (and incidentally, hourly) wages.

Quite apart from the fact that hourly wages thus fluctuate practically speaking in the same way as daily wages, Hanson offers us no explanation of why hourly wages should be constant and does not tell us which "institution" is implied. It would also be very queer if institutions, or, traditions were related to hourly wages, a wage form which is not and has never been used in rural Egypt. Labor is hired per day (or season, in the case of permanently employed hands), and it is significant that when the government in 1952 made an abortive attempt to institute statutory minimum wages in rural districts, it was a question of a minimum rate for *daily* wages.

If Hanson thus wants to challenge my interpretation of the *EPRA* data, he has to criticize the series for hours worked by actually employed laborers. This series is, indeed, open to some doubt. Strictly speaking it expresses the village sheik's opinion about hours to be worked at the daily wages of the week in question. He may be wrong and hours actually worked may differ from what has been agreed upon. I recall that when the *EPRA* was planned, I had expected the work time of employed laborers to fluctuate more than appears from the collected figures, not because of the fluctuations in demand for labor, but simply because the day is longer during the summer than during the winter. As in all old-fashioned agriculture the work day for hired labor in rural Egypt tends to be determined by the length of the day, and even at these southern latitudes the length of the day varies substantially during the year. Hence, my own expectation of both longer and greater fluctuation in hours of work per employed laborers.

There is no direct way of testing the reliability of the sheiks through the published material, but the following considerations may perhaps be taken as an indication that the sheiks knew what they were talking about. The statistical tables to the *EPRA* (I, Part I, Table 17, pp. 58–59) contain a frequency distribution of days by number of working hours, sex-age, and main occupation. Farm laborers are here separated out as a special group. Let us now define farm laborer days with one daily hour of work and more as "employed days." For male, employed farm laborer days, thus defined, the average number of hours work is 8.23, with a standard deviation of 1.78 hours, that is 22 percent measured upon the mean. For all males in the sample, including employed as well as unemployed days, the average number of hours is 6.33 with a standard deviation of 3.89 hours, that is 61 percent of the mean. Since our problem is to what extent the fluctuations of the average number of hours by all males (per 8 day period) can be assumed to express the fluctuations of the average number of hours for actually hired laborers, these percentage standard deviations are obviously relevant; they indicate a much larger variability in hours worked by all males than by employed laborers. However, the frequency distributions on which they are calculated are influenced by other circumstances than seasonality, for instance regional differences, farm and family size, etc. The measured standard deviations are thus larger than would follow from seasonality alone. And they are, of course, much larger than the standard deviations calculated on the figures behind the series shown in Figure 1, because on a given day in the year, all men do not work the same number of hours; the distributions of men by number of hours worked on a specific day are not available. If, however, regional differences[2] and farm and family size influence the distributions for farm laborers (which may have land, up to one-half feddan) and all men in the same way, the influence from

[2] I am not given access to the primary material any longer, but my recollection is that there were some differences between villages in working hours of hired labor.

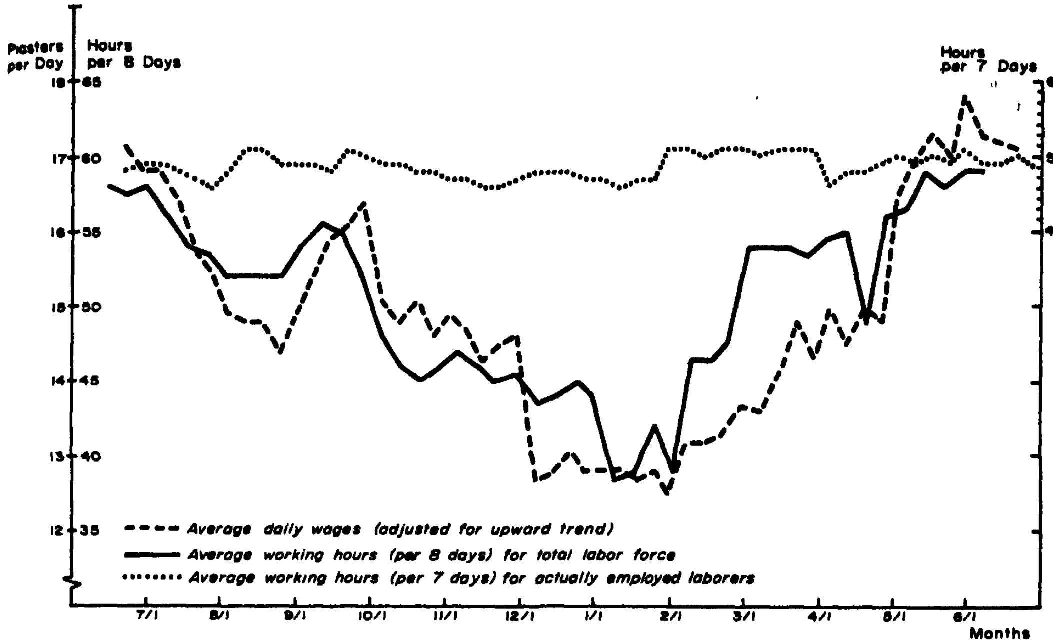


FIGURE 1

seasonality would also be in relation to the measured standard deviation. Now the average hours worked per 8 day period for all males is 50.91 and the standard deviation is 5.90 hours, that is 12 percent of the mean, which is about one-fifth of the percentage standard deviation of the frequency distribution of days by hours worked for all males, 61 percent. If the standard deviation of average hours per 8 day period for hired, actually employed laborers also were one-fifth of the standard deviation of days distributed by hours, which was 22 percent, one should expect the former standard deviation to be about 4 percent. The standard deviation of the sheik series is about 2 percent! And there is one factor which biases the measured standard deviations of days distributed by hours worked against the hypothesis of relatively constant working hours for hired laborers. As just mentioned, farm laborers may own or rent small plots of land on which they occasionally work; indeed, men in so-called landless agricultural families spent 31 and 27 percent of their work time on their own land (in small and big families, respectively). For laborers it seems likely, therefore, that days with very few hours have been spent on their own plot, and that days with very long hours have been spent partly as hired labor, partly on their own plot. If we could separate those hours which laborers have worked as hired labor, and this is what is relevant for our problem, we would probably find an even smaller percentage standard deviation for this group. No definite conclusions can be drawn on this basis, but it points at least in the direction of a much smaller variability during the year for average hours worked by actually hired labor than for all males in the sample.

Considering habits of life in Egyptian villages, I believe also now that about 8 hours effective work throughout the year may sound reasonable. Villagers rise and go to "bed" with the sun. At sunrise they get up and after a meal they are ready to walk or perhaps ride to the fields. Effective work may thus begin quite some time after sunrise. At midday there is always a long siesta and in the late afternoon, work has to stop early enough for laborers to reach their homes and get a meal before it gets dark. If we add that the midday siesta during the summer, when

temperatures in the shade may increase to 120 degrees Fahrenheit, has to be considerably longer than during the winter, I think we can understand why the effective work day for laborers tends to be roughly the same all the year round. If we shall speak of non-market determination of conditions of work for laborers in Egyptian agriculture, the work day for laborers may be the only case in point.

Hanson has three other objections to my analysis of rural wages. The first one is related to villages with missing information for the slack season. He argues on the one hand that to leave out these villages from the average implies a bias against the demand-supply theory (presumably because thereby the measured variation in the seasonal fluctuations diminishes?) while, on the other hand, missing information discloses missing employment opportunities. The first argument is wrong, and the second does not upset the demand-supply theory, which really, as I shall show in the next section, is the only way to explain why information sometimes should be missing in the slack season.

The second objection is more valid. An analysis of wages without regard to possible regional, and Hanson might have added, seasonal differences in costs of living may go wrong in that money wage differentials may simply reflect differences in costs of living. This argument is in principle correct in so far as regional differentials are concerned. I know of no systematic study of regional differences of costs-of-living for Egypt, but there may exist significant differences between the areas covered by the *EPRA*. Considering, however, the fact that a substantial number of items entering the budget of the lower income brackets in rural areas are subject to price controls with uniform prices, I doubt very much that cost-of-living differentials can explain much of the substantial regional wage differentials. But the problem ought to be studied and I may be wrong. Differences in cost of living cannot, on the other hand, have anything to do with the strong variations in the sex-age differentials, and the seasonal fluctuations in costs of living have in fact been studied. There is quite a natural tendency for the cost of living to have a seasonal trough during the summer when vegetables and fruits are in good supply; if anything, the seasonal fluctuations of real wages are thus slightly stronger than those of money wages (see Hansen, 1966, pp. 384–85).

Hanson's last objection is related to my suggestion of using the distribution of villages by average daily wages as a test of the two wage theories. To my eye the distributions seemed to support demand-supply theory rather than institutional wage theory, but partly because of the problem with cost of living, partly confronted with his high-powered statistical test, I agree with him that no definite conclusions can be drawn by this method, although I do not quite see why he insists upon leaving out the six lowest observations.

IV. Rural Employment and Demand-Supply Theory

I see no reason thus to give up the application of demand-supply theory of wages as expressed in the regressions on page 310 of my article. It should be stressed that the theory is at variance with the application of corresponding theory to developed countries (the Phillips curve). In the application to rural wages in Egypt, I assumed that wages adjust instantaneously, which means that the labor market is always cleared and no involuntary unemployment or underemployment can exist. The point is important because it implies that if at all we shall speak about unemployment and underemployment it must be in relation to some social norm. It may be useful to show how the theory works for slack seasons in villages where no laborer finds employment, and, consequently, it seems hard to insist that the market is cleared, and to deny that there is unemployment.

The following simple model of rural employment and wage formation is an adaptation of Chihiro Nakajima's model for family farms (see in particular, pp. 176 and 179–81). Nakajima considers only the case when the wage rate is given exogenously; our problem is the determination of both employment and wages. For the sake of simplification I

disregard work of farm family members outside the farms (small farmers working temporarily for big farmers, or for government) but the model can easily be extended to cover that case too.

Assume thus that we have farm families working on their own land, and landless laborers working for farm families. In Figure 2A, I have depicted the aggregate short run supply curve of farm families S^f, and landless laborers, S^l, respectively. Since the supply of farm family labor is used only for input on their own farm, I shall assume that the supply curve for farm family labor starts out from a lower level than that of landless laborers; this assumption is in line with the idea that self-employment is considered more respectable and/or enjoyable than work for others. Since we are considering the very short run, supply becomes zero at a positive wage rate. The total supply curve for labor is thus S. In Figure 2B we have the "open" or "net" labor market; labor supply in the open market is S^l. We make no distinction between hours and days, but this complication could easily be taken care of (see Sen).

Let then MP in Figure 2A depict the value of marginal productivity (at given farm product prices), falling with increased labor input.[8] The marginal productivity curve for a day or week at the peak season is MP_1. The corresponding net demand curve in the open market is D_1, which is the excess demand of farms, MP_1-S^f (horizontally). Utility maximization for the farm families requires that production is extended to the point where the market wage rate is equal to the value of marginal productivity, and that farm family labor is extended to the point where the farm family labor supply price is equal to the wage rate. At MP_1, total labor input becomes L_1, family labor input L_1^f and input of hired labor L_1^l. The wage rate is

[8] Let $Q_{j\tau}$ be the output of product j maturing for sale at time τ, and I_{it} be the input of factor i at time t, $t<\tau$. There are m products and n factors. The value of the marginal productivity of factor i at time t is then

$$MP_{it} = \sum_{\tau=t}^{\infty}\sum_{j=1}^{n} \frac{p_{j\tau}}{(1+r)^{\tau-t}} \cdot \frac{\partial Q_{j\tau}}{\partial I_{it}}$$

where $p_{j\tau}$ is the price of output j expected for time τ, r the interest rate per period (say, day), and the derivatives are calculated on the transformation function $F(Q_{j\tau};\, I_{it})=0$. The optimum condition is $W_{it}=MP_{it}$. Clearly MP will shift up and down during the year with the biological and climatic rhythms of agricultural production.

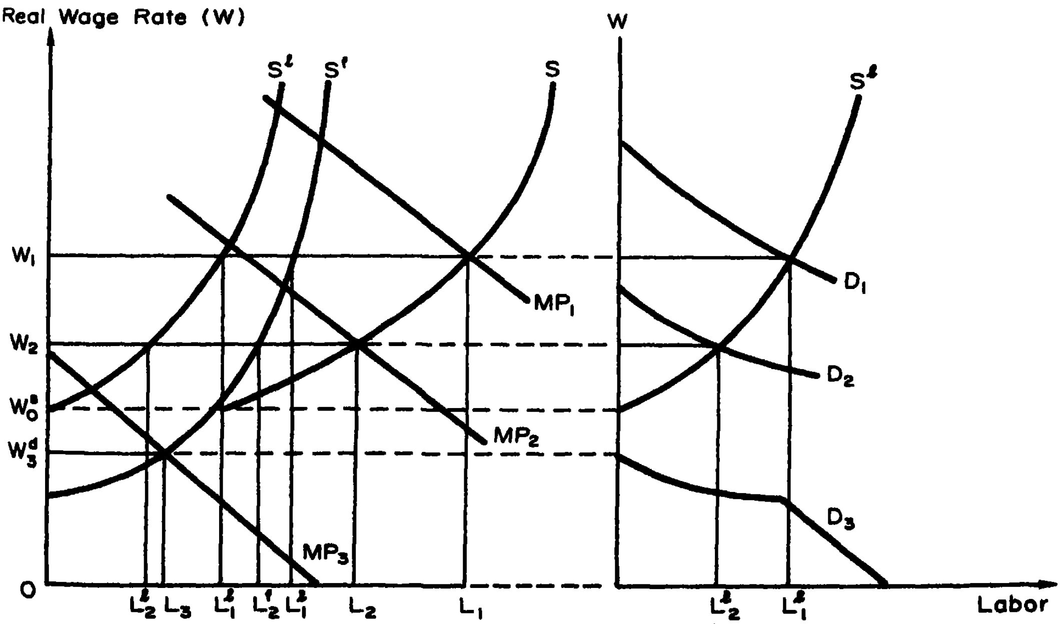


FIGURE 2A FIGURE 2B

W_1. At a less pronounced peak, the marginal productivity curve may be MP_2 and the net demand in the open market, D_2. Both family and hired employment will be lower, and so will the market wage rate which is now W_2. We note that no involuntary unemployment or underemployment have arisen; both farm families and laborers work at their optimum. Consider then a slack season with MP_3. Here farm family employment has fallen to L_3 and no hired labor is employed. Wages are *indeterminate*. The supply price for laborers at zero employment is W_0^s, while the demand price at L_3 is W_3^d. We have thus a situation where an equilibrium in the open labor market does not exist; no laborers are hired and there exists no market wage rate.[4] Marginal productivity is positive, but it could fall to zero without changing anything in the open market situation. If the MP curve passed through the origin (in which case not even family members would work), or the S^f-curve in the beginning coincides with the x-axis (in which case family members may do some work even at a very low marginal productivity curve), the supply price of laborers would still be W_0^s at a demand price equal to zero and the market wage rate would still be indeterminate. Even in this position we cannot determine the amount of involuntary unemployment, because there is no definite market wage rate at which supply could be measured.

The pattern described here corresponds closely to what actually happened in the wage survey. As the season became more slack, both wages and employment fell; and in some villages, before wages fell to zero, there were suddenly no wages to report; there was no active labor market, no hiring of labor. These are the villages with information about wages missing for the slack season. If there had existed an institutionally, or traditionally given wage rate, there is no reason why it should be impossible for the sheik to tell us what the wage rate is when there is no hiring of labor.

[4] What prevents a solution from existing is a discontinuity in the excess demand function. The proofs for the existence of a competitive equilibrium generally assume continuous excess demand functions.

V. Concluding Remarks

Since there is full agreement about the existence of some seasonal underemployment and unemployment, in particular in Upper Egypt, Hanson's concluding recommendations of counterseasonal employment policies can give rise to no disagreement either. It has always been recognized that the perennial cropping which now is being substituted for basin irrigation in parts of Upper Egypt will help to wipe out seasonal underemployment there. The main problem is how the electricity from the Asswan High Dam will be used. It seems now that priority will be given to village electrification which is probably a condition for creating counterseasonal employment in the villages and for mechanization of agriculture (repair shops). On the other hand, it should not be overlooked that to the same extent as seasonal underemployment of labor is diminished in this way, seasonal underemployment of capital may be created. It is not obvious what is best from a social point of view.

REFERENCES

B. Hansen, "Employment and Wages in Rural Egypt," *Amer. Econ. Rev.*, *59*, June 1969, 298–313.

———, "Marginal Productivity Wage Theory and Subsistence Wage Theory in Egyptian Agriculture," *J. Develop. Stud.*, *2*, July 1966, 367–407.

C. Nakajima, "Subsistence and Commercial Family Farms: Some Theoretical Models of Subjective Equilibrium," in C. R. Wharton, Jr., ed., *Subsistence Agriculture and Economic Development*, Chicago 1969, pp. 165–85.

A. K. Sen, "Peasants and Dualism with or without Surplus Labor," *J. Polit. Econ.*, Oct. 1966, *74*, 425–50.

Institute of National Planning, *Employment Problems in Rural Areas*, U.A.R.

(I) *Statistical Tables, The Labour Record Sample Survey*, Part I, Cairo, Dec. 1965; Part II, Cairo, Sept. 1966.

(II) *Report D on Wages, Income, and Consumption in Rural Areas*, prepared by B. Hansen, A. Sedki, and Y. Moustafa, Cairo, Dec. 1965.

ERRATA

The Determinants of U.S. Direct Investment in the E.E.C.

By Anthony E. Scaperlanda and Laurence J. Mauer

A number of errors occurred in Tables 1 and 2 of this article which appeared in the September 1969 *Review*, pages 558–68. The errors arose from the use of a regression program employing the single precision computational technique and in one instance, as the result of a scale change.

The corrected tables follow:

Table 1—Regression Equations for *U.S.* Direct Foreign Investments in the *E.E.C.*: Annual Data, 1952–1966[a]

	Regression Equation	$\bar{R}^2$	S_e	DW
(1.1)	$I=-426.707+5.520Y-58.860M+349.407\Delta M-8.099\Delta Y$	.952**	79.697	1.665**
	(5.171)** (.125) (.821) (.765)			
(1.2)	$I=-362.665+4.817Y-125.006M+337.617\Delta M-386.487G_1$	.950**	81.763	1.572
	(9.925)** (.248) (.774) (.239)			
(1.3)	$I=-364.682+4.764Y-188.711M+346.609M-.036G_2$	.950**	81.989	1.570
	(8.443)** (.349) (.702) (.041)			
(1.4)	$I=-463.272+4.899Y-.763M$	.955**	77.045	1.572*
	(12.129)** (.002)			
(1.5)	$I=-461.624+4.911Y+252.639\Delta M$	.957**	75.410	1.498**
	(17.783)** (.725)			
(1.6)	$I=-459.966+5.430Y-6.605\Delta Y$	.957**	75.451	1.720*
	(6.872)** (.716)			
(1.7)	$I=-436.758+4.899Y-293.815G_1$	.956**	76.889	1.599*
	(17.424)** (.221)			
(1.8)	$I=-471.167+4.928Y+.217G_2$	.956**	76.716	1.611*
	(16.762)** (.321)			
(1.9)	$I=-372.011+4.778Y-176.763M+337.201\Delta M$	.954**	78.180	1.572**
	(10.941)** (.406) (.809)			
(1.10)	$I=-518.935+5.616Y+118.506M-7.773\Delta Y$	.954**	78.510	1.670**
	(5.373)** (.288) (.746)			
(1.11)	$I=-454.159+4.938Y+50.536M-381.464G_1$	.952**	80.261	1.577**
	(10.953)** (.114) (.240)			
(1.12)	$I=-496.358+4.970Y+47.915M+.248G_2$	.952**	80.079	1.592**
	(10.548)** (.116) (.329)			
(1.13)	$I=-456.274+5.600Y+324.932\Delta M+8.533\Delta Y$	.957**	76.048	1.645**
	(6.842) (.901) (.894)			
(1.14)	$I=-410.177+4.912Y+289.040\Delta M-558.764G_1$	.954**	78.197	1.528**
	(17.150)** (.776) (.400)			
(1.15)	$I=-466.140+4.927Y+239.530\Delta M+.127G_2$	.954**	78.647	1.526**
	(16.347)** (.646) (.180)			
(1.16)	$I=-463.656+4.900Y$	.959**	74.022	1.571*
	(18.104)**			

[a] The numbers in parentheses are *t*-statistics; * and ** indicate significantly different from zero at 5 percent and one percent levels, respectively.

The symbol * and ** on the Durbin-Watson statistics indicates that the null hypothesis of residual independence cannot be rejected at the 5 percent or one percent level of significance. The tests of the *DW* statistics are based on the H. Theil–A. Nager testing procedure.

TABLE 2—REGRESSION EQUATIONS FOR *U.S.* DIRECT FOREIGN INVESTMENTS IN THE *E.E.C.*: ANNUAL DATA[a]

Regression Equation	$\bar{R}^2$	S_e	*DW*
1952–58:			
(2.1) $I = -245.359 + 3.294Y$ $(5.109)^{**}$	.807**	36.024	1.845
(2.2) $I = -339.707 + 2.736Y + 399.066M$ $(5.520)^{**}$ (2.553)	.908**	24.838	2.438
(2.3) $I = -243.567 + 3.303Y + 189.690\Delta M$ $(5.219)^{**}$ (1.092)	.814**	35.350	2.125
(2.4) $I = -245.903 + 3.060Y + 2.904\Delta Y$ $(2.954)^{**}$ $(.312)$	.765**	39.796	1.954
(2.5) $I = -260.521 + 3.299Y + 161.522G_1$ $(4.594)^{**}$ $(.192)$	.761*	40.092	1.905
(2.6) $I = -232.603 + 3.217Y - 0.153G_2$ $(4.417)^{**}$ $(.427)$	.769**	39.389	1.919
1959–66:			
(2.7) $I = -714.528 + 5.871Y$ $(11.064)^{**}$	.945**	75.069	2.629
(2.8) $I = -1069.750 + 6.470Y + 678.706M$ $(4.003)^{**}$ $(.396)$	.937**	80.971	2.493
(2.9) $I = -722.036 + 5.976Y + 1018.147\Delta M$ $(11.026)^{**}$ $(.990)$	.945**	75.196	2.550
(2.10) $I = -699.561 + 6.118Y + 3.696\Delta Y$ $(6.079)^{**}$ $(.300)$	.936**	81.503	2.747
(2.11) $I = -690.903 + 5.859Y - 225.577G_1$ $(9.820)^{**}$ $(.087)$	.935**	82.172	2.665
(2.12) $I = -697.549 + 5.835Y - 5.292G_2$ $(8.750)^{**}$ $(.107)$	.935**	82.139	2.642

[a] See note to Table 1 for explanation of * and **. Numbers in parentheses are *t*-statistics. None of the *DW* statistics were tested owing to the small number of degrees of freedom.

NOTES

Nominations for AEA Offices

The Electoral College on March 5 chose Kenneth J. Arrow as nominee for president-elect of the American Economic Association in the balloting to be held in the autumn of 1971. Other nominees (chosen by the nominating committee) are: for vice president (two to be elected), Ronald H. Coase, John G. Gurley, Hendrik S. Houthakker, and Arthur M. Okun; for member of the excutive committee (two to be elected), Gerard Debreu, Guy H. Orcutt, Joseph A. Pechman, and Melvin W. Reder.

Under a change in the bylaws reported in the *Papers and Proceedings* of this *Review*, May, 1971, additional candidates may be nominated by petition, delivered to the secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of president-elect and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the secretary at cost.

Announcement of Junior Dues Increase

At the annual meeting of the American Economic Association on December 29, 1970, those present voted to double the dues of all except junior members. A subsequent vote provided that the Executive Committee could raise the dues of junior members to the extent necessary to retain second class postal privileges.

The AEA counsel, Seymour J. Rubin, has ascertained that the dues of junior members need to be doubled for this purpose. The Executive Committee of the AEA voted, March 5, 1971, to increase the dues for junior members to ten dollars. An abstract of Mr. Rubin's opinion is given below:

> Part 132.228 of the "Post Office Regulations for Mail Classifications and Rates: Second Class" provides, essentially, that a publisher can charge a rate not less than 50 percent of its full subscription rate to a "premium group" of subscribers and still qualify for second class privileges, so long as the reduced rate charged is not less than the cost to the publisher of the subscription rate involved. Thus if your full subscription rate is $20, or some stated portion of the $20 membership fee you charge to regular members, you are allowed under the regulations to charge no less than half that rate ($10) to a premium group such as junior members, assuming the reduced rate is not less than your cost.

The American Academy of Political and Social Science recently held a conference, "Harmonizing Technological Developments and Social Policy in America." A monograph consisting of five principal papers, critiques on each, and proceedings of the conference may be obtained free by writing to the Academy, 3937 Chestnut Street, Philadelphia, Pennsylvania 19104.

The Committee on International Exchange of Persons announces availability of foreign economists for appointments in U.S. universities and colleges under provisions of the Fulbright-Hays Act for the academic year 1971–72.

Japan: Hiromi Ishigaki, Kaichi Shimura; Yugoslavia: Miroslav Petrovic, Branislav Soskic, Dragoljub Stojilkovic, Boris Tihi; Italy: Alessio Lokar, Alfredo Medio, Rino Ricci; Turkey: Turgent Var.

Inquiries should be addressed to: Miss Grace E. L. Haskins, Program Officer, Conference Board of Associated Research Councils, 2101 Constitution Avenue, Washington, D.C. 20418 or via telephone—Area Code 202, 961-1948.

The National Science Foundation will sponsor a summer institute for junior college instructors of economics, August 2 27, 1971. The topics will include contemporary economic analysis and the economics of education. Director: H. Robert Heller, economics department, University of California at Los Angeles, Los Angeles, California 90024.

The National Institute of Social and Behavioral Science will hold sessions for contributed papers at the 138th annual meeting of the American Association for the Advancement of Science, December 26–31, 1971, in Philadelphia. Economists who would like to give a paper at these sessions are invited to forward titles and abstracts of 300 words by August 25 to Donald P. Bay, National Institute of Social and Behavioral Science, 863 Benjamin Franklin Station, Washington, D.C. 20044.

Suggested topics for papers might include in whole or in part the economics of selected wage-price controls; a national income policy; the Federal full-employment budget in theory and practice; percentage state income tax deductions from federal income taxes as an optimum form of revenue-sharing; environmental and welfare economics; some aspects of scope, degree, and intensity in the redistribution of income in the United States; reciprocal trade agreements questions and foreign dumping as current problems in the U.S. balance of payments; employment and the status of free trade theory in a high inflation economy; automation, computers, and management science as instruments of social and economic change in the U.S.S.R.; foreign economic competition in postwar Southeast Asia and problems in the protection of economic nationalism; general international investment, trade, and exchange questions; economic development theory; and subjects in interdisciplinary areas.

The proceedings of a research conference, held Sept. 11–12, 1970 by the Inter-University Committee on Urban Economics, are now available. The papers cover

urban land and housing markets, economic consequences of air pollution, poverty, public goods theory, and related policies. Orders enclosing $3.00 payable to the University of Chicago should be addressed to Professor G. S. Tolley, department of economics, 1126 E. 59th Street, Chicago, Illinois 60637.

A two-day meeting will be held at Oklahoma State University, Stillwater, on the relationships between computer science and statistics. The keynote speaker will be H. O. Hartley of Texas A&M University. Workshops will be conducted in five areas: TIME SERIES AND STOCHASTIC PROCESSES: Emanuel Parzen, State University of New York, Buffalo, chairman: DECISION SCIENCES: Dennis Grawoig, Georgia State University, Atlanta, chairman: COMPUMETRICS: Robert Gordon, University of California, Irvine, chairman, COMPUTER SCIENCE AND STATISTICS IN HIGHER EDUCATION: J. L. Folks, Oklahoma State University and Ron Mohler, University of Oklahoma, co-chairmen; COMPUTER SCIENCE AND STATISTICS IN THE EXTRACTIVE INDUSTRIES.

Persons having papers in any of these five areas (up to about eight pages) are invited to submit them to the conference chairman, Dr. Mitchell O. Locks, Oklahoma State University, Stillwater, Oklahoma 74074.

The American Academy of Political and Social Science, 3937 Chestnut Street, Philadelphia, Pennsylvania 19104, announces that it will send free upon request, copies of a study "Design for International Relations Research: Scope, Theory, Methods, and Relevance."

Deaths

A. B. Andarawewa, economics branch, Canada Department of Agriculture, Ottawa, fall 1970.

Frederick A. Bradford, professor emeritus, Lehigh University, Jan. 22, 1971.

Elwyn Cady, Kansas City, Missouri, Dec. 15, 1970.

Herbert B. Dorau, professor emeritus of economics, School of Commerce, New York University, Jan. 17, 1971.

Eberhard M. Fels, University of Munich, Germany, spring 1970.

Paul Gekker, Board of Governors of the Federal Reserve System, Jan. 16, 1971.

Everett B. Hurt, assistant professor, Scarborough College, University of Toronto, Feb. 14, 1971.

Jerzy F. Karcz, professor of economics, chairman, Russian area studies, University of California at Santa Barbara; general editor, *Eastern European Economics*; Santa Barbara, Dec. 10, 1970.

Doris G. Phillips, professor of economics, California State College, Fullerton, Jan. 14, 1971.

Robert Rockafellow, professor of economics, University of Rhode Island, Dec. 16, 1970.

Erich Schneider, Germany, Dec. 5, 1970.

Gilbert Stonesifer, professor and chairman, department of economics and business administration, Mount Union College, Oct. 1970.

Clair Wilcox, professor of political economy emeritus, Swarthmore College, Dec. 31, 1970.

Retirements

Roland B. Eutsler, professor of economics, University of Florida, Oct. 1, 1970

Morris E. Garnsey, department of economics, University of Colorado, fall 1971.

Robert W. Harbeson, professor of economics, University of Illinois, Urbana, Aug. 1971.

Ben W. Lewis, International Division, The Ford Foundation, Aug. 31, 1971.

Ralph Sherman, Ohio State University, Feb. 1, 1971.

Carl S. Shoup, professor of political economy, Columbia University, June 1971.

Elbridge Sibley, Social Science Research Council, Dec. 31, 1970.

N. Arnold Tolles, Cornell University: professor of economics, New York State University at Geneseo.

Leland S. Van Scoyoc, professor of economics, Bowling Green State University, Dec. 1970.

Paul Webbink, Social Science Research Council, Dec. 31, 1970.

Visiting Foreign Scholars

H. G. Brennan, Australia National University: visiting foreign scholar, department of economics, Dalhousie University, 1971–72.

Bienvenido Delantar, Trinity College of Quezon City: visiting assistant professor in economics, Trinity College, fall 1970–Jan. 1972.

Kenneth A. Leslie, University of the West Indies, Kingston, Jamaica: Food Research Institute, Stanford University, 1970–71.

Broadus Mitchell, Hofstra University: visiting professor of economics, University of Rhode Island, spring 1971.

Luigi Pasinetti, Cambridge University: visiting professor, Columbia University, fall 1971.

Joan Robinson, Cambridge University: visiting professor of economics, University of Waterloo, fall term 1971.

Marcel M. Tardos, Institute for Economics and Market Research, Budapest, Hungary: visiting lecturer, department of economics, University of Michigan, winter 1971.

Alfred Zauberman, London School of Economics and Political Science: visiting lecturer in economics, University of California, Santa Barbara, winter, spring 1971.

Promotions

M. William Belovicz: associate professor of economics, University of Massachusetts, Feb. 1, 1971.

Nicholas G. Bohatiuk: professor of economics, Le Moyne College.

Lawrence A. Boland: associate professor, department of economics and commerce, Simon Fraser University.

E. Gerald Corrigan: chief, domestic research division, Federal Reserve Bank of New York.

Richard V. Cotter: professor in managerial sciences, University of Nevada, July 1, 1971.

Ralph C. d'Arge: associate professor of economics, University of California, Riverside.

Michael Edelstein: assistant professor of economics, Columbia University.

Samuel M. Ehrenhalt: deputy regional director, Bureau of Labor Statistics.

Stanley L. Engerman: professor of economics, University of Rochester, Sept. 1, 1971.

James Esmay: associate professor, San Fernando Valley State College.

Frederick Finch: associate professor of economics, University of Massachusetts, Feb. 1, 1971.

Raymond S. Franklin: associate professor of economics, Queens College of the City University of New York.

Martin Geisel: associate professor of economics, Carnegie-Mellon University, Sept. 1, 1971.

Louis Geller: associate professor, Queens College of the City University of New York.

Harry I. Greenfield: professor of economics, Queens College of the City University of New York.

Everett E. Hagen: director for center of international studies, Massachusetts Institute of Technology.

William Haller, Jr.: professor of economics, University of Rhode Island, July 1, 1971.

Harry J. Halley: deputy controller and program director for management systems, Board of Governors of the Federal Reserve System, Jan. 1971.

John J. Hooker: professor, department of economics and finance, University of Texas at El Paso.

Peter E. Kennedy: associate professor, department of economics and commerce, Simon Fraser University.

Leonard Lapidus: assistant vice president, Federal Reserve Bank of New York

Larry J. Larsen: professor of economics, University of Nevada, July 1, 1971.

Lester B. Lave: professor of economics, Carnegie-Mellon University, Sept. 1, 1971.

L. C. Ledebur: associate professor of economics, Denison University.

Will Lyons: professor of economics, Franklin and Marshall College.

Donald Mathieson: assistant professor of economics, Columbia University.

Kent Monroe: associate professor of economics, University of Massachusetts, Feb. 1, 1971.

Woo H. Nam: associate professor of economics, San Diego State College.

Peter Passell: assistant professor of economics, Columbia University.

Malvika Patel: associate professor, department of economics, San Fernando State College.

John D. Patrick: chief, balance of payments division, Federal Reserve Bank of New York.

James L. Pierce: associate adviser, division of research and statistics, Board of Governors of the Federal Reserve System, Jan. 1971.

A. Marshall Puckett; adviser, research and statistics function, Federal Reserve Bank of New York.

Fredric Q. Raines: associate professor of economics, Washington University, St. Louis, July 1, 1970.

Robert Rivers: professor of economics, University of Massachusetts, Feb. 1, 1971.

Dick L. Rottman: professor of managerial sciences, University of Nevada, July 1, 1971.

Gregory C. Schmid: chief, foreign research division, Federal Reserve Bank of New York.

James Starkey: assistant professor of economics, University of Rhode Island, July 1, 1971.

Rudolf Thunberg: manager, domestic research department, Federal Reserve Bank of New York.

Sheila Tschinkel: special assistant, securities department, Federal Reserve Bank of New York.

William E. Whitesell: associate professor of economics, Franklin and Marshall College.

Parker Worthing: associate professor of economics, University of Massachusetts, Feb. 1, 1971.

Administrative Appointments

Alec P. Alexander: academic assistant to chancellor, University of California, Santa Barbara, Jan. 1971.

Wallace N. Atherton: chairman, department of economics, California State College at Long Beach, Feb. 1, 1971.

M. E. Bond: director, bureau of business and economic research, Arizona State University, July 1971.

Flournoy A. Coles, Jr.: director, center for black economic development, Fisk University.

Josephine M. Corrigan: assistant to the dean, St. John's University, Sept. 1970.

Richard S. Cowan: chairman, department of economics and business administration, Waynesbury College.

Simeon J. Crowther: associate chairman, department of economics, California State College at Long Beach, Feb. 1, 1971.

Ward S. Curran: director of institutional development, Trinity College, 1971–73.

Dennis J. Dugan: chairman, department of economics, University of Notre Dame, Jan. 1, 1971.

Melvin A. Eggers: vice chancellor for academic affairs and provost, Syracuse University.

Joseph A. Giacalone: assistant dean, St. John's University, Sept. 1970.

Harold M. Hochman: director of studies in urban public finance, Urban Institute.

Kenneth C. Kehrer: acting chairman, department of economics and business administration, Fisk University.

William H. Leahy: director of undergraduate studies, University of Notre Dame, Jan. 1, 1971.

Gordon W. McKinley: senior vice president, economics and financial planning, McGraw-Hill, Inc.

Hugh S. Norton: chairman, department of economics, University of South Carolina, fall 1970.

James E. Price: chairman, economics department, Syracuse University.

G. Ray Prigge: chairman, department of economics and business administration, Mount Union College.

Clarence A. Reed: manager of developmental re-

search, marketing research department, The Coca-Cola Company.

M. Selim: director of center for economic education, College of St. Thomas.

Roger M. Skurski: director of graduate studies, University of Notre Dame, Jan. 1, 1971.

T. Arthur Smith, U. S. Army director of cost analysis: chairman, Defense Economic Analysis Council.

Richard J. Solie: head, department of economics, University of Alaska, Sept. 1970.

Peter O. Steiner: chairman, department of ceonomics, University of Michigan.

Vito Tanzi: chairman, department of economics, The American University.

J. Vanderkamp: chairman, department of economics, College of Social Science, University of Guelph, July 1, 1971.

New Appointments

I. Robert Andrews, Bowling Green State University: associate professor, department of economics and commerce, Simon Fraser University.

Harold Barnett, Federal Reserve Bank of Boston: instructor, department of economics, University of Rhode Island, fall 1970.

Randall Bartlett: assistant professor, department of economics, Williams College.

William J. Baumol: professor of economics, New York University.

H. Prescott Beighley: economic research unit, division of research, F.D.I.C., Washington

Sanford V. Berg: assistant professor of economics, University of Florida, 1971–72.

Rudolf M. Binnewies, University of California at Berkeley: instructor, department of economics, Vassar College.

Harold A. Black: assistant professor of economics, University of Florida, 1971–72.

Richard Boisvert: assistant professor, department of economics, Cornell University.

Myles G. Boylan, Jr.: instructor of economics, Case Western Reserve University.

William A. Brock, University of Rochester: visiting associate professor of economics, University of Chicago, 1971–72.

Karl Brunner: professor of economics, joint appointment, department of economics and Graduate School of Management, University of Rochester.

John Burge: visting assistant professor, department of economics, Mercer University, Sept. 1970.

Charles Cathcart: assistant professor of economics, Pennsylvania State University, Sept. 1971.

Don M. Chaffee: assistant professor of economics, University of South Carolina, fall 1970.

Divakar J. F. Chandran: assistant professor of economics, University of Alaska, Sept. 1970.

L. Duane Chapman, University of Tennessee: assistant professor, department of economics, Cornell University.

Henry E. Cole, Johns Hopkins University: assistant professor of economics, Tulane University.

Flournoy A. Coles, Jr., Fisk University: professor of economics, Graduate School of Management, Vanderbilt University.

Steven G. Darr: instructor, department of economics and business administration, Mount Union College.

Martin H. David, University of Wisconsin: professor of economics, University of Michigan.

J. Ronnie Davis: associate professor of economics, University of Florida, 1971–72.

Robert M. Deaver: lecturer in economics, Arizona State University, spring 1971.

Rudiger Dornbusch: assistant professor of economics, University of Rochester, Sept. 1971.

Ernest G. Ellingson: assistant professor of economics, Ball State University.

Dennis Ellis, Wayne State University: assistant professor, department of economics, Southern Illinois University.

Asim Erdilek: instructor of economics, Case Western Reserve University.

Edward C. Ettin: assistant adviser, division of research and statistics, Board of Governors of the Federal Reserve System, Jan. 1971.

Robert Falconer: economist, financial statistics division, Federal Reserve Bank of New York.

Terry A. Ferrar: assistant professor of economics, Pennsylvania State University, Sept. 1971.

Ray W. Fowler: professor of buisness, department of economics, Walla Walla College, July 1970.

Mark W. Frankena: lecturer, department of economics, University of Western Ontario.

Richard J. Gelson: economist, financial statistics division, Federal Reserve Bank of New York.

Gary G. Gilbert: economic research unit, division of research, F.D.I.C., Washington.

Peter Gregory, University of Minnesota: professor of economics, University of New Mexico.

Klaus D. Grimm, educational testing service, Princeton: research associate, department of public social services, county of Marin, California.

Herbert G. Grubel, University of Pennsylvania: professor, department of economics and commerce, Simon Fraser University.

Peter E. Gunther: assistant professor of economics, Mount Allison University.

Demos Hadjiyanis: associate professor of economics, College of St. Thomas.

Peter Heller, Harvard University: assistant professor, department of economics, University of Michigan.

John Hill: economist, domestic research division, Federal Reserve Bank of New York.

A. G. Holtmann, University of Wisconsin; professor of economics, State University of New York at Binghamton.

Yutaka Horiba: assistant professor of economics, Tulane University.

George C. Hoyt, University of Iowa: professor, department of economics and commerce, Simon Fraser University.

David B. Humphrey: associate professor of economics, Tulane University.

Thomas O. Jones, Jr.: professor of economics and management, U.S. Coast Guard Academy, fall 1970.

Harry Kelejian: professor of economics, New York University.

Clifford R. Kern, Harvard University: assistant professor of economics, State University of New York at Binghamton, Sept. 1971.

Karen I. Kidder: economist, banking studies department, Federal Reserve Bank of New York.

Young Kim: assistant professor, department of economics, Northern Illinois University.

Tetteh A. Kofi: visiting professor, Food Research Institute, Stanford University, 1970–71.

L. Emil Kreider: assistant professor of economics, Beloit College, fall 1970.

Peter J. Kuch: lecturer, department of economics, University of Western Ontario.

Robert Kuller, University of Kansas: assistant professor, department of economics, University of Wyoming.

Leonard Kunin: associate professor of economics, San Jose State College, fall 1971.

Roslyn Kunin: visiting assistant professor, department of economics and commerce, Simon Fraser University.

Abba P. Lerner, University of California, Berkeley: professor of economics, Queens College of the City University of New York, 1971–72.

Roger D. Little, University of Houston: assistant professor, economics department, U.S. Naval Academy, Aug. 1970.

Peter C. Lin: assistant professor of economics, University of Alaska, Sept. 1970.

Ronald G. Lorentson: acting assistant professor of agricultural economics and acting assistant economist, experiment station, University of California, Berkely, fall 1970.

Jerolyn R. Lyle, U.S. Equal Employment Opportunity Commission: lecturer in economic theory, Smith College, spring 1971.

Michael D. McCarthy: associate professor, department of economics, Case Western Reserve University, Sept. 1971.

Thomas O. McCoy: assistant professor, department of economics, Williams College.

Donald G. McFetridge: lecturer, department of economics, University of Western Ontario.

Peter F. M. McLoughlin, University of California, Santa Clara: visiting professor of economics, department of political economy, University of Toronto, Sept. 1970.

James M. Mangum, Oklahoma State University: associate professor of economics, Louisiana Tech University, Sept. 1970.

Harry Mapp: assistant professor, department of economics, Cornell University.

Laurence H. Meyer: assistant professor of economics, Washington University, July 1, 1969.

J. Robert Moore, Millersville State College: assistant professor of business administration, University of Wisconsin, Feb. 1, 1971.

Yasuo Murata: professor, department of economics, Dalhousie University, July 1, 1971–June 1973.

Michael Mussa: assistant professor of economics, University of Rochester, Sept. 1971.

Soren T. Nielsen: instructor, department of economics and commerce, Simon Fraser University.

Richard B. Norgaard: acting assistant professor of agricultural economics and acting assistant economist, experiment station, University of California, Berkeley, fall 1970.

Ronald L. Oaxaca: visiting lecturer, department of economics, University of Western Ontario.

Thomas R. O'Connor, State University of New York at Buffalo: instructor, department of economics, Wabash College.

E. Odgers Olsen, University of North Carolina: assistant professor of economics, State University of New York at Binghamton, Sept. 1971.

John P. Palmer: lecturer, department of economics, University of Western Ontario.

Alan Pankratz: assistant professor of economics, DePauw University, 1970–71.

Clifford H. Patrick: economic research unit, division of research, F.D.I.C., Washington.

Prasanta K. Pattanaik: associate professor, department of economics, University of Western Ontario.

Nicholas Perna: economist, domestic research division, Federal Reserve Bank of New York.

Bruce Petersen, National Bureau of Economic Research: instructor, department of economics, University of Wyoming.

Manferd O. Peterson: economic research unit, division of research, F.D.I.C., Washington.

John T. Pickerill: assistant professor of economics, Ball State University.

James P. Quirk: professor of economics, division of the humanities and social sciences, California Institute of Technology, fall 1971.

Robert Rice: assistant professor of economics, University of Hawaii.

Sergio Roca: instructor, department of economics, Adelphi University.

Richard K. Rudel: assistant professor, economics department, South Dakota State University, Feb. 1, 1971.

William R. Russell, University of Kentucky: professor of economics, Southern Methodist University.

Stephen R. Sacks: assistant professor of economics, University of Connecticut.

David T. Scheffman: visiting lecturer, department of economics, University of Western Ontario.

Peter J. Schmidt: assistant professor, department of economics, Wayne State University, Jan. 1, 1971.

Robert N. Schoeplein, University of Connecticut: associate professor, department of economics and institute of government and public affairs, University of Illinois, Urbana.

William L. Scott, Jr.: instructor, economics and business, Mercer University, Sept. 1970.

David E. Seckler, University of California, Berkeley: professor, department of economics, Colorado State University, Mar. 25, 1971.

Lawrence C. Smith, Mississippi College: professor of economics, Louisiana Tech University, June 1970.

Robert S. Smith: assistant professor of economics, University of Connecticut.

Norman H. Starler, South Dakota State University: assistant professor of economics, State University of New York College at Fredonia.

William Stull; assistant professor of economics, Swarthmore College.

John Swirles, University of British Columbia: visiting associate professor, department of economics and commerce, Simon Fraser University.

Leslie Szeplaki: assistant professor, department of economics, Northern Illinois University.

J. Ernest Tanner, University of Western Ontario: associate professor of economics, Tulane University.

Reed Taylor: department of agricultural economics and rural sociology, Ohio State University, Jan. 1, 1971.

Thomas H. Tietenberg: assistant professor, department of economics, Williams College.

Ronald L. Tinnermeier, North Carolina State University: associate professor, department of economics, Colorado State University 1971–72.

Tore Tjersland: professor, department of economics, Colorado State University, 1971–72.

Lynn Turgeon: visiting professor, department of economics, San Jose State College, fall 1971.

Susan Wachter: lecturer in economics, department of economics, Bryn Mawr College, 1971–72.

William Walsh: assistant professor of economics, College of St. Thomas.

David Weinberg, University of California at Berkeley: instructor, department of economics, Vassar College.

Roger S. White: assistant professor of economics, University of Connecticut.

Ronald Wilder, Vanderbilt University: assistant professor, department of economics, University of South Carolina, fall 1970.

Robert S. Woodward: lecturer, department of economics, University of Western Ontario.

Gregory J. L. Yi: assistant professor of economics, West Virginia University.

Leaves for Special Appointments

Martin Bronfenbrenner, Carnegie-Mellon University: visiting professor of economics, Aoyama-Cakuin University, Tokyo, fall 1970.

William E. Gibson, University of California at Lo Angeles: Brookings Economic Policy Fellowship, economic research unit, division of research, F.D.I.C. Washington.

Philip D. Grub, School of Government and Busines Administration, George Washington University: visiting professor of international marketing, economic faculty, University of Helsinki, spring semester 1971.

Woo Sik Ki, West Virgina University: research department, Korea Exchange Bank.

Arlyn J. Larson, Arizona State University: legislativ research assistant, Arizona Legislature, spring 1971 summer 1972.

Huntley G. Manhertz, University of Pittsburgh Institute of Social and Economic Research, Universit of the West Indies.

David C. Warner, Wayne State University: senio economist, Health and Hospital Corporation of Ne York City.

Resignations

Leslie Aspin, Marquette University: Representative 1st Congressional District, Wisconsin.

Robert W. Baguley, University of Western Ontario June 1971.

Thomas J. Beirne, University of Nevada, May 1971

John A. Dominick, Louisiana Tech University: University of Arkansas, Aug. 1970.

Peter F. Donnelly, Federal Reserve Bank of New York: Public Utility Commissioner, Akron, Ohio.

Peter B. Kenen, Columbia University: Princeton University, June 1971.

Patricia Keough, Federal Reserve Bank of New York: New York University School of Economics.

Jacques Melitz, Tulane University.

Thomas F. Wilson, Federal Reserve Bank of New York: Butler University.

Miscellaneous

Arthur D. Lynn, Jr., Ohio State University: associate editor, *National Tax Journal*.

The American Economic Review

VOL. LXI, NO. 3 PART 2 JUNE, 1971

SUPPLEMENT

SURVEYS OF NATIONAL ECONOMIC POLICY ISSUES AND POLICY RESEARCH

Price $1.00

Foreword

The two surveys of national economic policy issues and policy research published in this Supplement to *The American Economic Review* are the third in a series of studies of economics in foreign countries, commissioned by the Publications Committee of the American Economic Association and under the editorship of Harry G. Johnson. The previous series, edited by George W. Hildebrand, sought to acquaint Anglophone economists with the significant developments in economics since the Second World War in the national literatures of the major non–English–speaking countries. This series, by contrast, is directed at economic policy issues, and the controversies and research to which they have given rise, in countries selected for one or more of three reasons: the intrinsic interest of the policy issues, the relevance of the policy issues in the United States, and the interest of the countries themselves as areas of involvement of American foreign economic policy with which U.S. economists are likely to become concerned.

HARRY G. JOHNSON

Mexican Economic Policy in the Post-War Period: The Views of Mexican Economists

By Leopoldo Solís*

TABLE OF CONTENTS

Mexican Economic Policy in the Post–War Period: The Views of Mexican Economists

By LEOPOLDO SOLÍS*

Introduction

In Latin America, where inflation and economic stagnation seem to be the common rule, the case of Mexico is almost unique. In the past twenty-five years it has experienced sustained economic growth at a rate of more than 6 percent a year, almost double the rate of population growth. Output per capita has reached 560 dollars and Mexico ranks twelfth among the economies of the free world in terms of its gross national product—28 billion dollars. In the process, the country underwent a period of development with inflation, from the end of the Great Depression until 1956, followed by a period of growth with relative price stability that has persisted to the present.

Many of the obstacles to economic development have been overcome through the creation of an extensive base of social overhead capital and the establishment of institutional mechanisms which make for greater mobility of the factors of production, an essential condition for steady growth. Nevertheless, these factors are insufficient to guarantee the continuation of the process of growth. As economic progress takes place, it requires new social overhead capital projects, institutional changes, and the creation of instruments and incentives within economic policy which will enlarge and improve the orientation of investment, as well as the development of skills and the allocation of labor.

The first section presents an introductory description of the outstanding features of the economic behavior and of the nature of economic policy in Mexico since World War II, focusing on topics such as the agricultural sector, the industrial sector, the distribution of income, the financial system, the public finances, and the balance of payments. The purpose of this material is to give the reader a general background with which to appraise the subsequent sections, where opinions on domestic economic policy sustained by Mexican economists, and comments on the same topics by other experts whose views have been influential during the same period, are presented. The second and third sections summarize the views on certain aspects of agricultural and industrial supply and their elasticity. The fourth section examines aspects of allocation, the distribution of income, and consumption and marketing. The fifth and sixth sections describe monetary and fiscal policy respectively. The seventh section summarizes points of view which have been expressed in relation to the foreign sector and the eighth relates to economic development. The ninth section briefly explains the economic planning approaches that have been suggested in Mexico. In the tenth and final section, the author expresses his views on the opinions presented, and on some areas not covered in

* Head, Department of Economic Studies, Bank of Mexico, and Professor of Economics, El Colegio de Mexico.

revious sections, as well as on the activies of Mexican economists.

I. *Background*

The recent economic history of Mexico falls into two clearly defined stages. The first of these, characterized by development with inflation, covers the period from the end of World War II until 1956, with annual price increases of 11 percent, government deficits financed with currency issues, foreign exchange instability —the peso was devaluated in 1948 and again in 1954—and a rapid agricultural development. In fact, agricultural output had a higher rate of growth than that of the gross national product (6.2 percent) and brought about a considerable expansion in agricultural exports and foreign trade in general. In short, this was a period of externally–oriented growth.

The second stage, however, has been inwardly-oriented, based on substitution for imports of manufactured goods, the proportion of which decreased from 8 percent in the first period to 5 percent in the second. A relative stagnation has also become evident in the agricultural sector, the rate of growth of which fell below that of the gross national product, which has remained 6.2 percent. The exchange rate of the peso has been maintained at 12.50 to the dollar. Prices have only risen 1.9 percent on the average in recent years—less than in the United States, with which Mexico effects 70 percent of its foreign trade.

As a result of the diminished vitality of the agricultural sector, the rate of increase of agricultural exports has declined considerable, and in general receipts of foreign exchange have increased at a slower rate than expenditures abroad. This has produced a large deficit in the current account of the balance of payments and an increase in the foreign debt. The slowing down of investment in social overhead capital and in agricultural research and development, as well as the distortion of cost structure resulting from the effects of protectionism on the prices of industrial inputs, have had an adverse effect on the competitive position of agricultural exports and constitute a factor in the recent slowdown in growth of merchandise exports. On the other hand, the parity prices for goods destined for domestic consumption have been raised in order to stimulate their output. (The parity prices are set by the official regulating agency for agricultural commodities, *Compañia National de Subsistencias Populares.*)

One characteristic of the period of price stability has been the growth of domestic savings, which has played an important role in the accelerated expansion of the banking system and, in fact, in a general financial boom of unusual magnitude, brought about by factors such as transfers from the non-financial market to the institutional sector. The process has facilitated the placement of government securities by the central bank among the financial intermediaries—especially the non–monetary ones—so that government deficits no longer represent a mere creation of money by the central bank, as they did in the first period.

The Agricultural Sector

The agricultural sector has provided many of the outstanding features of the economy of the country. This sector absorbs 50 percent of the labor force and contributes less than 20 percent of the value added in the production process. These facts in themselves express the problems related to the distribution of income and the productivity of the labor force. The problems are much more pronounced when the dual nature of Mexican Agriculture is taken into account. There is a modern agricultural sector that can compete internationally in terms of effi-

ciency and productivity and which generates about 65 percent of the agricultural output while employing only 36 percent of the agricultural labor force. The traditional sector, which continues to have a backward technology and a low productivity, employs the larger part of the agricultural labor force and produces ten percent of the value-added in agriculture.

Despite the problems mentioned, the agricultural sector has been strategic in the process of development. During the first stage of recent economic development 1946–1956, the average annual rate of growth of the gross national product, at constant prices, was 6.2 percent. This growth was characterized by the rapid development of the agricultural sector, which grew at a rate of 7.5 percent a year, that is, faster than the growth of domestic demand for agricultural commodities due to the increase in both population and real income. Thus, it was possible to improve the payments position of the country by substituting for imports of agricultural commodities and creating agricultural surpluses for the export markets. Foreign sales of agricultural products, especially those coming from the new irrigated areas, constituted the prime factor in the increase of exports, offsetting the decline in earnings on merchandise account due to the stagnation of mineral exports, which had been evident since the end of the Great Depression.

The rapid growth of the agricultural sector was fostered by the land reform movement and investment in public works to promote agriculture, which resulted in a better distribution of agricultural output and the creation of a commercial, modern and efficient agriculture. Moreover, research led to the application of new, superior techniques which lent a greater flexibility to agricultural supply.[1] Aside from the expansion of the area under cultivation, the irrigated areas increased more than proportionally, thus reducing the importance of random climatic conditions. Similarly, agricultural research led to the development of new and superior varieties of seeds and to the propagation of modern farming techniques, including the use of fertilizers and insecticides, especially in the irrigated zones. As a result, there was a more flexible supply, as well as increases in total agricultural output, in the yields per hectare, and in output per man. The dual nature of Mexican agriculture thus became more accentuated. On the one hand, a commercial agriculture emerged in the irrigated zones, with high productivity and a capacity to absorb technological change, which has increased the yields, uses modern inputs with a high import content and is not labor intensive. In contrast, there is the traditional subsistence agriculture, largely oriented to the production of corn, making intensive use of labor with low productivity, unreceptive to new techniques and new crops, and in which population growth represents an increasing pressure on the resources.

Since 1950, there has been a new trend in the allocation of public funds for investment. Works for the promotion of agriculture, which had had priority within public investment since the 1934–1940 Cardenas regime (when the redistribution of land gained momentum) became less important and, instead, emphasis was placed on public investment for industrial development. The effect of public works in agriculture, after a period of transition during which the large irrigation projects went into operation, consisted of an increase in the rate of growth of agricultural output. The new policy had the converse effect—a decline of public investment in

[1] The available figures show that between 1941 and 1951 there was an average annual increase of 2.9 percent in the area under cultivation and of 1.0 percent in the yields per hectare. The average increases between 1953 and 1962 were 1.3 percent in the area under cultivation and 1.9 percent in the yields.

agriculture, including education and research, that could not be offset with increased yields or the (now reduced) expansion of irrigated lands. In fact, the rate of growth of agricultural output decreased to 3.8 percent.

The slower rate of growth of the agricultural sector was lower than the rate of increase in the domestic demand for its products. Consequently, between 1956 and 1967, there were no increases in agricultural surpluses for the export market.[2] Moreover, parity prices for commodities destined for the domestic market were raised by the regulating agency for agricultural prices, to the detriment of the relative price of export commodities, which did not rise in world markets. Besides, as a result of industrial protectionism, there were increases in the prices of the industrial inputs most frequently used in export products, such as cotton, and a subsequent loss in their profitability. Thus, both agricultural output and exports had low rates of growth. In the period 1955–1964, total exports expanded at an average annual rate of 4 percent, while those of agricultural products grew only 1.4 percent a year.

The pattern of behavior of Mexican agriculture poses serious problems to the economic development of the country. On the one hand, the modern agricultural sector has not increased output sufficiently to meet domestic demand, hence restricting the exports and therefore the import capacity of the country. On the other hand, the limited technological development of the traditional sector has made the absorption of the growing agricultural labor force increasingly difficult, reducing the levels of per capita incomes and bringing about a migration toward the urban centers, which in turn had generated cheap labor for industry and the service sector, exerted pressure on public utilities, and made it necessary to increase investment of social interest.

[2] In the period 1956–1968, the rate of population growth was 3.4 percent a year and that of product per capita was 2.7 percent. Both factors combined to make the rate of growth of the demand for agricultural commodities about 5.5 percent.

Industrial Development

Even prior to the drastic changes in agricultural production and in the yields per hectare, which permitted the growth of exports and secured the necessary foreign exchange for imports of machinery and equipment, industrial development has been the primary goal of economic development policy since World War II. Industrial firms were protected from international competition by means of tariffs and restrictive import permits. Tax incentives were created and, with the support of financial policy, an increasing volume of funds was channeled into industrial investment frequently by means of loans from official banks and international organizations at lower interest rates than those prevailing in the domestic market. Regulations were established so that private credit institutions would increase their industrial portfolios. Similarly, foreign firms, previously oriented toward raw material exports and public utilities, began to manufacture for the domestic market. The expansion of the industrial base strengthened the urban market for agricultural commodities in such a way that domestic demand supplemented exports as a stimulus to agricultural output.

These factors explain the 6.2 percent annual growth rate of manufacturing between 1940 and 1953—a rate almost equal to that of total output in the same period. In the years 1953–1965, the rate of growth rose to 8.3 percent a year, far above that of the gross national product in the same period. The development of the steel industry, the metal products manufacturing industry and the chemical industry spurred the growth of the whole sector. Yet some of the traditional indus-

tries—such as textiles, apparel and shoes—grew at rates similar to that of agricultural output. On the whole, the importance of industry was enhanced, as evidenced by the increase of its relative share of gross national product from 30.8 percent in 1950 to 35.7 percent in 1965. The relative share of manufacturing industry in the same aggregate rose from 20.8 percent to 25.7 percent. Nevertheless, this information in itself does not reveal the important internal structural changes within manufacturing. Other industrial activities developed in accordance with special conditions that affected their pattern of behavior.

The electric power and petroleum industries, stimulated by the government effort to strengthen the productive social overhead capital, expanded more or less at the same rate as manufacturing. Their average annual rates of growth in the first period (1940–1953) were 6.5 and 6.3 percent, respectively. In the second period (1953-1965), like manufacturing, they accelerated their rate of growth to 9.7 percent and 7.6 percent respectively.

The industrial evolution of Mexico between 1950 and 1965 was achieved by means of the accelerated production of capital goods and intermediate goods. In effect, ranking industries in accordance with the use of their products, there has been a decline in the importance within the industrial structure of those industries oriented toward the production of consumer goods. Consumer goods represented 72.2 percent of total manufactures in 1950 and 43.3 in 1965; the annual rates of growth for the fifteen-year period under consideration were 5.6 percent for consumer goods industries and 11.1 percent for producer goods industries. Industrial development in the same period strengthened the production of intermediate and capital goods. The dependence on imports of inputs necessary for the maintenance of equipment not produced domestically did not become more pronounced or more rigid.

The high level of tariff protection and quantitative restrictions on imports have meant high differentials of domestic over foreign prices, as well as high rates of effective protection which pose serious problems for industrial development. Similarly, although it is highly protected, industry has an unutilized capacity, uses capital-intensive processes, and has a limited absorption of labor.

Distribution of Income

In the last three decades, however, while Mexico has undergone a process of sustained economic development that has permitted output per capita to increase from 75 dollars in 1940 to 560 dollars in 1968, there have been indications that the process of development itself has engendered inequalities in the income received by the various segments of the population. The gains in output per capita must thus be qualified by adverse changes in the distribution of income.

In most countries, factor payments to labor represent a stable proportion of income. In Mexico, however, this proportion has been subject to wide fluctuations. Salaries and wages represented 29.2 percent of total income in 1940, but only 23.8 percent in 1950. In the period of growth with price stability, the trend was reversed, and the share of salaries and wages increased to almost the same proportion it had represented in 1940. The decline in the relative share of labor in comparison to that of capital throughout the decade of the 1940s can be explained by the rapid changes in prices and the greater use of labor during the war. This, together with the small possibility of increasing the stock of capital, meant that output growth could be obtained only through additional work shifts, thus increasing profits and

the relative share of capital in the distribution.

The problem becomes patent when the data on income distribution by productive sectors and income brackets is taken into consideration. Agriculture, which is the activity that still absorbs the highest proportion of the labor force—50 percent—has an average level of income far below that of other sectors (less than 25 percent of the total in 1967). In addition, the distribution of income within the sector itself is very unequal, for 53 percent of the people engaged in agriculture receive only 23 percent of agricultural income. Inequality has been accentuated because commercial agriculture, with few units, grows more rapidly than the traditional, subsistence agricultural sector. In the period of rapid agricultural growth the relative (and even the absolute) position of low-income families self-employed in agriculture did not improve, while there was a rise in the income of landowners, agricultural entrepreneurs and farm hands of the advanced segment of agriculture in the irrigated areas.

Although initially the agrarian reform brought about a redistribution of income, its effect in absolute terms has been very small. On the other hand, public investment outlays on works for economic development, by making the modern sector of the economy more dynamic (especially in agriculture), have fostered a greater polarization of the economic situation without improving the lot of families in the lower income brackets, particularly in the agricultural sector.

In the industrial sector, the situation is much more equitable, for almost half of the people employed by it receive about 42 percent of the income. The mobility of rural labor toward the urban centers and its absorption by industry have not been sufficient to correct the inequitable distribution of income. Since the relative share of the agricultural sector within national income decreases faster than the proportion of the labor force engaged in agriculture, the distribution of income not only continues to be inequitable, but tends to become more so as economic development takes place.

In general, the role of the public sector in the distribution of income has not been significant. While it is difficult to evaluate the effect on the different strata of expenditures on social works, which have grown considerably, it can be assumed that they are not very important as a redistribution mechanism. Expenditures on education, which account for a large part of these outlays, have a greater effect on those who pursue their studies beyond elementary school, who, for the most part, do not come from the rural areas or the lower income groups. A similar situation is true of public health services, which primarily benefit the urban centers.

Some economic policy measures, such as the concession of tariff protection and quantitative controls on imports, tend to raise the price of consumer goods and thus distort the distribution of income. Moreover, these measures give rise to disparities between domestic and world prices, which reduce the real income of the consumers of protected goods. The associated transfers of real income favor profits and other returns to property, because the conditions of a surplus of labor imply a practically constant wage rate.

The Financial System

In the last ten years, the Mexican financial system has experienced a sustained and steady development, as a result of the growing volume of savings. In former years, the financial system was greatly limited in its operations by the uncertainty resulting from inflation and the low real rate of interest, for there were fixed limits on the nominal interest on bank lia-

bilities and bonds. Voluntary domestic saving, measured by bank liabilities in national currency, remained fairly constant in relation to national income in the inflationary period. More recently, however, its rate of growth has surpassed that of the gross national product at current prices. The strengthening of the Mexican financial system has permitted a more efficient investment of savings. Also, its rapid growth has given the authorities a greater weight in the money market in terms of attracting and channeling funds through the financial system. Thus, some economic sectors that had not been considered attractive credit risks, despite being key sectors for development, can now obtain credit through the mediation of the monetary authorities, who have resorted to diverse mechanisms in order to facilitate their access to credit. The natural resistance of the commercial banking sector, which normally wants to invest in activities that constitute a low risk or of which it has previous knowledge, has thus been overcome.

It has therefore been possible also to transfer the voluntary savings of the community to the public sector and to finance government deficits without jeopardizing monetary stability. The funds transfered in this way represent up to two-fifths of the liabilities of the banking system. Consequently, the growth of the banking system itself has been conditioned by the annual increase in total liabilities. This type of financing does not have an inflationary effect, for it uses the real savings of the population and channels them toward non-monetary intermediaries. The liabilities of the Mexican banking system in domestic currency have increased at an annual rate of 18 percent over the past ten years. The growth rate, however, was slightly lower in 1967 and 1968, perhaps as a result of the decrease in the volume of savings outside the banking system, which during the latter years of economic stability constituted a significant part of the flow of funds toward the banking system.

Domestic savings channeled through private credit institutions have increased at an annual average rate of 21 percent since 1960; those channeled through the public institutions have done so at a rate of 22 percent. Within the liabilities of the private institutions, the growth of investment banks ("financieras") has been outstanding, for they have had a mean annual growth of 33 percent since 1960 and an even more marked increase in the past two years. This does not hold true for other private credit institutions, especially commercial banks, which have suffered a decline in their relative importance.

Although the instruments used to attract funds have become more diversified, their principal characteristic is still their liquidity, for even new financial certificates are convertible on demand, despite a fixed maturity date. The high liquidity of the system is one of its most vulnerable points, for in times of uncertainty the structure of holdings can be altered quickly to other real assets or to foreign currency, as occurred during the devaluation of 1954.

Public Finance

Mexican fiscal policy has had the following main objectives: the determination of a desirable composition of government spending in consumption and investment; the extension of services and assistance lent by the State to the community, spreading educational, medical and social welfare services to the population; the use of tax incentives to foster saving and private investment; making the tax burden more comprehensive by means of income taxes; and finally, the use of subsidies and tax exemptions to stimulate industrial development. With respect to the balance of payments, fiscal policy has consisted of creating incentives for industrialization by

means of tariffs, exemptions and subsidies, in order to obtain substitution of domestically-produced goods for imports.

The main obstacles confronting Mexican fiscal policy have been: a persistent tendency toward budgetary disequilibrium and the problems of financing it in a non-inflationary way because of the weakness of the securities market; a concentration of the distribution of income and the problems of taxing the higher income groups on remuneration from various sources hard to discover; and the problem of taxing members of *ejidos* and cooperatives, which predominate in certain sectors, for political reasons.

The breakdown of the ordinary revenues of the Federal Government provides an idea of the tax structure of the country and its sources of funds. In 1967, total Government revenues amounted to 39.4 billion pesos—12 percent of gross national product. Of this total, income taxes represented 28.9 percent, while taxes on consumption accounted for 21.9 percent of the total, and taxes on income received from services and property of the State amounted to 4.2 percent of the total.

Another important source of Federal funds has been the public debt, although the rudimentary nature of the money and capital markets limits the volume of domestic debt. For this reason, especially since the slowdown of exports in the period of stability, foreign loans have been sought by the Government from international organizations and private institutions. In this way, domestic savings have been supplemented by foreign credits, with the result that there have been larger volumes of investment and equilibrium in the financing of development.

The recent income tax reforms represent an attempt to obtain larger volumes of the income of the community by consolidating the former cedular tax system —based on different rates depending on the income source—into an overall income tax rate. The reform of the tax system was begun in 1961. The taxable base was enlarged in order to include types of income which formerly had been exempt, such as returns on fixed-income securities, capital gains, and income from rentals of real estate. Similarly, a supplementary tax rate was created for accumulated income (as a precedent to the tax on total personal income), which offset to some extent the lack of progressiveness in the rates. In order to foster reinvestment, undistributed profits were exempted from taxes, provided that they were used for reinvestment. Another stimulus for investment was a scheme for accelerated depreciation.

The Federal Tax Bureau (Registro Federal de Causantes) was created in order to control those subject to taxation. Since the former system resulted in a low level of government revenue, and because it was inequitable (taxpayers within the same brackets paying different rates according to the source of their income), a change was made in 1965 to tax total income of both individuals and firms. The tax rate for firms with an income of more than 500,000 pesos is 42 percent. The level of taxable income was reduced for certain incomes received by credit institutions, insurance firms, and bond companies.

The recent trend of Mexican fiscal policy has been to reduce the dependence of the tax system on indirect taxes, which are regressive by nature, and to revise taxes with low yields which both make for high collection costs and induce a high degree of tax evasion.

However, there are still structural flaws. Studies show that it is still necessary to reform taxation on income and spending. Excise taxes include a number of items which have low yields due to high administration costs associated with the establishment of various levels of rates with no thought of integration; and often the rates are graded pyramidally, which is a draw-

back because of overlapping taxes. In 1966, further changes were made to integrate the budgets of government agencies and enterprises with that of the Federal Government. Similarly, all of the applications for credit of these institutions now have to be approved by the Ministry of Finance.

Public spending has been used to create the social overhead capital framework of the country. This was done initially by lending support to agricultural development and communications and later, in the recent period of stability, by doing the same for industrial development and works of social interest. In 1966, 42.9 percent of total public spending was allocated to economic development, 29.5 percent to investment and social welfare, 7.1 percent to the armed forces, 11 percent to administration costs, and 9.5 percent to servicing the public debt.

Balance of Payments

After the exchange instability and the boom in exports during the inflationary period had ebbed, there was a decline in the growth of exports, posing serious problems for the maintenance of the exchange rate and the growth of the national product. The problem was partially solved by means of financial policy, especially that concerned with maintaining the growth of the import capacity. This was achieved mainly through foreign credits, which, in turn, permitted a high rate of growth of the gross national product combined with price stability.

The deficit in the current account of the balance of payments has become more pronounced in recent years.[8] Between 1963 and 1968, expenditures abroad at current prices increased at an average annual rate of 10.6 percent, while foreign exchange earnings increased 6.7 percent. The result has been a larger deficit in current account, which in turn has meant an increase in foreign indebtedness. It is significant, however, that there was a change in the structure of expenditures abroad, for merchandise imports increased at a lower rate than overall expenditures and their relative importance declined, while that of expenditures by Mexican tourists abroad increased, as did the remittances on direct foreign investment and the interest payments on foreign debt of the public sector, which became twice as important.

Imports have shown a tendency to decrease as a proportion of the gross national product. This has been especially true in the period of price stability, of rapid industrial development, conditioned by import substitution.

High tariff protection has been a recent development, and very marked in the period of stability. The effect of tariffs is stronger than is generally thought. The tariff structure reveals low duties on raw materials, machinery and basic foodstuffs and high import duties on manufactured consumer goods.

In Mexico, the balance of payments is a secondary consideration in the formulation of trade policy. The strongest argument, since the decade of the fifties, has been the protection of incipient industries, as development became increasingly oriented toward the domestic market. During this period, the substitution for imports of durable consumer and intermediate goods gained momentum. But this process was very intensive as early as 1950.

The items which have contributed to the growth of foreign exchange earnings between 1963 and 1967 are income receipts from transactions along the border with the United States, tourism, agricul-

[8] In the period between 1950 and 1968, with the exception of 1950 and 1955—years immediately following devaluations of the peso—the deficit on current account averaged 258.6 million dollars annually. In 1962, the deficit was 156.4 million dollars. The average for the last four years under consideration amounted to 446.1 million dollars.

tural commodities, manufactured goods and mineral products. Merchandise exports had the lowest rate of growth—4.2 percent. Foreign earnings from exports of services grew at a rate of 9.6 percent. The latter rate was abetted by tourism, which had an annual growth of 14.6 percent, making it the most dynamic item in the balance of payments, as well as by foreign exchange earnings border transactions with an annual growth of 7.5 percent. These trends have been reflected in the structure of income in current account. Merchandise exports fell to 50.2 percent of total exports in 1967, while income from tourism increased to 16.5 percent of the total.

As a result of the slow growth of merchandise exports, income on current account of the balance of payments available to pay for imports after subtracting payments to foreign factors (profits of foreign investment and servicing foreign credits) grew at an annual average rate of 6.1 percent between 1963 and 1967—less than the 8.5 percent rate of growth of imports.

II. *The Agricultural Sector*

In discussing policy related to the agricultural sector, it is necessary to distinguish between policy concerning land tenure and policy concerning agricultural production. The differentiation is chronological, as well. Land tenure policy was predominant during the period of intensive redistribution of land, starting during the Lazaro Cardenas regime (1934–1940). Later, interest gradually began to center on the problems of agricultural production stemming from agricultural development and the new pattern of land tenure.

Land Tenure

What were the results of the agrarian reform? Under what conditions is the land tilled in Mexico? Agrarian reform changed the structure of the ownership of land and, consequently, wrought changes in the rural social structure itself. Its principal goal was social justice—the distribution of land among the peasants who work it. For this very reason, it lacked an economic criterion as a guideline for the distribution of land to establish productive units of an efficient size. Because of this, there is still much discussion about the optimal size of plots. However, according to E. Flores (1967), it involves a good deal of ideological discussion and few reliable studies of the economic feasibilities.

With the land reform came the *ejido* system,[4] giving a communal solution to the agrarian problem. This system was created mainly on the basis of the lands expropriated from the former *haciendas* (large land holdings), which, by the same token, brought about the disintegration of the latifundia. The disappearance of the latifundia had two immediate consequences: the destruction of a great number of productive units and a subsequent decline in the productivity of others, and the divorce of economic and political power from the traditional rule of class (R. Stavenhagen, 1966, pp. 468–469).

A form of small private ownership of the land emerged together with the *ejido* system. In 1960, barely 25 percent of the agricultural labor force was made up of *ejidatarios*, which reveals that the majority of the rural population lives under a system of private enterprise. There are also owners of medium sized properties and family farms, lying somewhere in between the small and the large landowners.

Most opinions concur that the *ejido* is in crisis or stagnant and is not an institution than can provide a satisfactory solution to the agrarian problem. Stavenhagen points out that the *ejido* system of land

[4] On this *ejido* system the land is owned by peasants, who can work it and bequeath it, but can neither sell it nor rent it.

tenure has created several problems as the result of communal production under the conditions of a capitalist economy. For example, the collective *ejidos* have been infiltrated by interests and political forces which have thwarted them. Within the *ejido* system, the main problems are a) the smallness of the individual plots or *minifundios*, b) the uncertainty of tenure, and c) the scarcity of credits for the *ejidatarios*. These problems, which stem from the smallness of individual plots, hinder the formation of capital as well as an increase in productivity, and maintain the technological level of the rural sector stagnant. R. Fernandez y Fernandez (1957) points out some of the symptoms of the situation—the sale of plots to strangers, the marked government paternalism, the rigidity of the institution (or the fact that it merely represents a subsidiary form of land holding, which is not very efficient).

A problem to which other authors refer is that of the lack of flexibility in the performance of the *ejido*. The more efficient *ejidatarios* cannot enlarge their holdings of land—which, by law, they cannot rent —nor can they displace the less efficient. The problem has become more acute because of demographic growth and the lack of new lands in areas where the agrarian reform already has been effected. Nevertheless, there are many instances of *ejidatarios* who lease their land illegally in order to become salaried workers.

But how did these conditions arise? The consensus of opinion points to the very mechanism of the agrarian reform and the organization with which the *ejidos* were actually endowed as the principal factors. Fernandez, M. Hinojosa Ortiz (1961), and Stavenhagen make the observation that in the distribution of land, recipients were not screened nor were they organized technically or institutionally. The corps of technicians in charge of implementing distribution was insufficient, both in number and in qualifications, and generally lived in the cities. On the other hand, there have been numerous shortcomings in its operation, and its organization on the basis of assemblies and commissaryships has only been an additional problem. There has also been a dispersion of the jurisdiction of the authorities which has led to confusion among farmers (Hinojosa). In conclusion, the *ejido* was a system that was conducive to the substitution of labor for capital (M. A. Duran, 1968), in an environment which is still overpopulated (Fernandez, 1957). As Stavenhagen remarks, besides the fact that the majority of the *ejidatarios* were given extremely small plots of inferior quality, the lower productivity of the *ejido* can also be attributed to the irrigation policy, which accorded greater benefits to the owners of private plots. The *ejido* has lost its initial momentum and, at present, it is being thwarted by the private agricultural sector (Fernandez, 1957), which has the most productive form of landholding (Stavenhagen).

The proposed solutions to the problem contain many viewpoints in common. In this respect, Fernandez y Fernandez (1957) speaks for many authors. He points out the need to select the new *ejidatarios* who will receive land to be distributed; to determine which *ejidos*, and in what fields, will adopt the cooperative form (with regard to the existing cooperatives there are some discrepancies). For some, this form of organization is the best course to follow and its radius of action should be enlarged (Duran, 1961; Fernandez, 1954); for others it is merely an abstruse imitation with a complex structure (H. Flores dela Peña and A. Ferrer, 1951). Fernandez also stresses the need to increase the size of the plots held by every *ejidatario* to the size of family farm plots; to make it possible for the *ejidatario* to

purchase land; to chastise the *ejidatario* who does not work his plot; to organize them, seeking a structure to foster internal cooperation; to maintain a system of equal plot sizes by regions, since it is practically the only feasible one (however, others differ on this and would prefer that size depend on the economic conditions of the land); and to insist that government intervention be limited to the administration of justice and technical assistance.

Within the private sector of agriculture, the principal problem is the small private plot and the conditions around it, which are getting worse. The majority of the owners of small plots live on a subsistence level. Like their *ejidatario* peers, they are characterized by low productivity and a limited technical level, besides which they are still in the hands of local money-lenders. Many of them have to look for work as migrant farm laborers in the United States, or as temporary laborers on large farms, or migrate to cities. On the other hand, commercial farmers are very efficient and with highly mechanized enterprises. In the small plots and the *ejidos*, there is a surplus of labor which does not favor the formation of capital. A great number of them produce for home consumption. The small plot hinders technical progress. The incidence of the role of technological change is reflected in that part of output earmarked for home consumption (F. Rosenzweig Hernandez, 1963). This is the reason why greater productivity can be obtained on larger properties and agricultural progress is more marked in the areas under irrigation (Flores dela Peña, 1953; G. Lira Porragas, 1964).

Similarly, regarding the commercialization of agricultural products, there are monopolistic conditions conducive to price-fixing that hamper the formation of capital in agricultural units. It is frequently pointed out that the middlemen make the largest profits, for they buy from small, unorganized farmers—to whom they often have lent money—and later sell as monopolist to the regional market (Rosenzweig, 1963; Stavenhagen, 1966).

For Rosenzweig as well as for Stavenhagen, the small plots constitute the main problem of the agricultural sector today, above all because they impede the formation of capital for lack of stimuli. Moreover, the situation is taking a turn for the worse. No coherent policy has been formulated to solve the problem and it cannot evolve "naturally," due to the structure of the markets and the lack of technical education.

The solution for the small private and *ejido* plots, as seen by Fernandez (1954) and Rosenzweig (1963), is to integrate the fractionized property into plots of an economically efficient size. At the same time, the small farmers, who constitute a weak and scattered group, must count on greater government protection. Another writer, Flores de la Peña (1954), argues that if the system of extensive exploitation is continued, there will also be the drawback of not solving the population problem. Because of this, small farms of the European type—intensively worked—should be established as a measure to solve the problem of land tenure as well as that of overpopulation.

One cannot overlook the advance in the productivity of agriculture. The social overhead capital works executed by the government permitted substantial progress in various areas, especially in the northern part of the country. Similarly, new means of communications, namely highways, allowed easy access to markets.

There are two positions regarding what is almost always called the *latifundium*, that is, the commercial exploitation of relatively large extensions of land. One of them, upheld by Hinojosa (1961), con-

siders that the Agrarian Reform put an end to the *latifundium* in Mexico. The other, sustained by Flores de la Peña, Fernandez and Stavenhagen, among others, is that the existence of new *latifundia*—modern, commercial agriculture on large properties—is a fact in the country today. According to Flores de la Peña, Mexico after 30 years of Agrarian Reform, is a country of large landholders. In Stavenhagen's opinion, this situation is recent and for that reason he speaks of neo-latifundia. He claims that the new landowners are often very efficient, highly mechanized agricultural enterprises and that they represent the negation of the very ideals of the Agrarian Reform (1966). Flores de la Peña (1954) and Lira (1964), point out that the concentration of property can be found particularly in the areas under irrigation, and they are in agreement with Stavenhagen in attributing the prevalent form of *latifundia,* of recent origin, to the way land was distributed in newly–irrigated regions. In any case, all agree on the need to abolish the *latifundium.* Fernandez, (1960) favors forbidding it by law and attacking clandestine or de facto *latifundia.*

Underemployment and Colonization

The phenomenon of rural underemployment is universally recognized. It is also a factor that plays an important role in recommendations for development policy, as will be shown later. Opinions on this topic supplement each other. For Flores (1964, Ch. XX), the principal problem arising from this situation is the waste of human resources. One of the reasons for disguised unemployment is the traditional outlook of the population and its concentration in certain regions. Flores de la Peña feels that rural underemployment is a problem that has yet to be solved and because of this proposes the establishment of small farms with an intensive use of labor as the predominant mode of agricultural production.

A number of solutions have been proffered with respect to the problem of underemployment. The main one is a measure that would be feasible only in the long run: the absorption of the surplus rural population by secondary and tertiary activities. This, however, would depend on the growth of these activities. Flores (1964), Flores de la Peña (1958), and Rosenzweig (1963), agree with this viewpoint, but there are others. For example, according to Flores de la Peña (1954), the rural population should be frozen at six million people, as of 1960, who would be guarenteed work all year round. He also makes other recommendations, such as the displacement of groups of farmers towards the northern part of the country or the Gulf region, which contain the most productive lands. Finally, he is of the opinion that underemployment will not be solved by the modernization and mechanization of agriculture nor by the distribution of lands (1958, p. 378).

Colonization has been carried out very slowly (Fernandez, 1960) and is increasingly costly, since the best lands have already been distributed (Rosenzweig, 1963). For Stavenhagen, colonization has been important in recent years: because of it, the increasing importance of internal migration in the country is due to factors which are not directly related to agrarian reform (1966, p. 468). Flores de la Peña considers colonization preferable to the distribution of lands under cultivation, for reasons of capital formation and the improvement of overall efficiency. Fernandez advises the acceleration of this process in view of the overpopulation on the *ejidos.*

Agricultural Policy

Capital Formation. On the whole, there has been a limited capitalization of the

private segment of the rural sector. There is general agreement on this, even though there are differences of opinion regarding its causes. Rather than being antagonistic, however, the opinions tend to supplement one another. Fernandez (1960) believes that the reason lies in the lack of incentives, emanating, as shown above, from the pattern of behavior of the *ejidos*. On the other hand, Rosenzweig holds that the reason can be found fundamentally in the monopsonistic practices of the markets, as well as the deterioration and forsaken condition of capital equipment on some of the land (1963). To this, other experts add the insufficiency of credit.

The evaluation of public investment is different. The works executed by the government, such as communications and irrigation projects, are considered essential factors in the determination of the increase in productivity experienced in recent decades. Thus, Flores (1959) influenced by J. Schumpeter, speaks of the "strategic innovation" of these projects, in the sense of the effect that they had on the entire productive structure and its process of transformation. At the same time, they fostered the relocation of agriculture, displacing it toward the northern and northwestern parts of the country, as well as stimulating migration toward new farming areas.

In view of the foregoing, the solutions proffered are obvious: all insist on more intensive investment, private as well as public. For some, like Duran, Flores and Rosenzweig, planning must be introduced into agriculture. Others consider that the irrigation and communications policy must be continued, while the credit facilities and technical assistance granted by the government are expanded (Flores de la Peña, 1958). Rosenzweig adds that there should be a more extensive use of chemical inputs in agriculture (1963). Finally, Hinojosa manifests some dissention by suggesting that the imbalance between large and small irrigation be corrected (1961). In this, he sides with Stavenhagen, who states that the irrigation works have basically benefitted not the ejidatarios and small farmers, but the large landowners (S. de la Peña, 1963).

Commodities and Markets. Flores links the relocation of crops with the importance acquired by cotton, coffee and henequen, which were organized as commercial agricultural production oriented toward foreign markets. Yet, another phenomenon has emerged: these commodities are being displaced by fruit and vegetable crops because of the higher productivity of the latter two (1959). Regarding wheat, there are two contradictory opinions. Flores maintains that this commodity has not reached the position attained by the others, despite a favorable domestic price and credit policy. However, Lira (1964) believes that it is the only commodity within Mexican agriculture that has good yields.

The opinions about markets can be separated into those pertaining to foreign markets and those referring to domestic markets. Most experts have doubts about the stability of the foreign markets. According to Duran, their instability is always latent (1961). Similarly, Lira affirms that difficulties are still encountered in placing the output of the northwestern region in foreign markets (1964). There is, furthermore, agreement on the flagging vitality of the domestic market, probably a reflection of Engel's law. Fernandez (1954) and Flores de la Peña (1954) attribute the lag in the consumption of the agricultural masses to the deterioration of their terms of trade with the industrial sector. Moreover, Flores de la Peña refers to the lack of an economic policy on basic commodities. On the other hand, Rosenzweig (1963) and Flores de la Peña (1954), point out the prevalence of monop-

sonistic situations in the domestic market. The first of these authors further indicates that the low level of domestic consumption is a result of low productivity and that agricultural demand is weak and easily satisfied.

All of these authors are in favor of enlarging the domestic market. Rosenzweig offers some concrete solutions with regard to the objectives for economic policy, such as the organization of the market and producers, the elimination of middlemen, improving the commercialization of agriculture and abolishing monopolistic situations. Fernandez (1954) stated in the early 1950s that the State should intervene marginally in fixing prices by means of regulating inventories.

Technology and Credit. The paltry investment in agriculture is related to the low level of technology (Flores, Rosenzweig, Duran and P. Padilla). Yet, there have been regional differences in the use of technology and the gains have not reached all areas evenly, according to Hinojosa (1961).

The problem of credit has been dealt with extensively, so that it is possible to infer that it manifests itself as a phenomenon of real importance in the rural sector, and that it is closely related to the insufficient investment and the market situation. All of these authors agree that investment is insufficient in terms of the "needs" of producers, but there are different opinions about the relationship between cause and effect. According to some (Fernandez, Duran, Flores and Flores de la Peña), the lack of credit is a determining factor in the low productivity of the rural sector, since it is insufficient, poorly timed, and centralized. Others (Rosenzweig and especially Stavenhagen) consider that the lack of credit is a result of the mode of agricultural production—namely small farms and the market situation—which makes the credit risk involved very high, so that the private banking system does not meet the needs of the sector. Finally, it should be mentioned that credit to *ejidos,* in practice, only benefits a small proportion of the *ejidatarios* (Stavenhagen, 1966) and in many instances, there has been malversation on the part of those in charge of *ejido* credit funds (Hinojosa, 1961). Those authors who have dealt with the subject point out the need for improvement. Fernandez and Flores specifically refer to the need to dencentralize the credit system. Lira emphasizes that it is necessary to extend the terms of credit and to allocate it to annual crops and fruticulture. Rosenzweig does not believe in half-way measures—it is true that improvement is necessary, but it should be conditioned by other measures to increase productivity (1963).

General Analysis of Agriculture

A general appraisal of the agricultural sector embraces two outstanding aspects—the result of the agrarian reform and the role of agriculture within the context of economic growth.

The agrarian reform had positive and negative aspects. According to Flores (1959), its most important contribution was essentially the redistribution of wealth, income, and power. With this, it gave the initial impetus to economic development, propitiated the mobility of factors of production, and altered the pattern of land use. Moreover, it spread its influence to other economic sectors. Fernandez (1954), and Rosenzweig (1963), consider it the reason for public investment in agriculture, that is, in social overhead capital projects such as irrigation, communications, and electricity, which later helped to increase productivity. At the same time, it was not without flaws, stemming from the absence of a well-defined program and its fundamentally political motivation, as well as economic limitations and human

shortcomings (Rosenzweig, 1963; Stavenhagen, 1966). This has led Stavenhagen to the conclusion that the reform has not yet been completed, for it has only attained its goal of redistributing the land and has inadvertently permitted the emergence of large landowners, with an efficient production closely tied to commercial agriculture.

The role of agriculture within economic growth has not been less important. Flores de la Peña (1954), considers that it has been essential in maintaining an external equilibrium—an opinion shared by other economists. Nevertheless, the sector has not been favored, as the same author points out, for development has meant a decrease in the real income of the rural sectors. The contraction of the effective demand for agricultural output has only accentuated this phenemenon (Flores de la Peña, 1958). Other authors find the limitations on agricultural growth in the behavior of the factors of production—the scarcity of capital and the surplus of labor (Padilla and Rosenzweig). Finally, Lira (1964), leaning toward a concept to be examined later, feels that the disequilibria of the sector are of a structural nature and that their solution requires economic planning in agriculture.

How can the ideas examined thus far be summarized? What stands out is that, despite its real achievements and impact on the present economic expansion, the agrarian reform has left some situations that require change, so that the agricultural productive system can reach a level of efficiency that would be compatible with the needs of the agricultural masses. Demographic pressure in the countryside is substantial, which merely represents an additional element that impedes the formulation of solutions.

One final appraisal is necessary: there is a scarcity of works that analyze livestock breeding and the recent stagnation in the production and exports of agricultural commodities. In speaking of the rural question, almost all of the authors are referring to topics related to agriculture, while the livestock aspect receives only marginal attention, such as recommending its increase in new areas or fostering consumption in the domestic market. It is well known that there is a small consumption of animal proteins and a high income–elasticity in the demand for these products.

III. *The Industrial Sector*

Background

In the past twenty years, industrialization has been interpreted as the very essence of economic development in Mexico. In the vainglory of nationalism, industrial self–sufficiency represents both the economic independence of the country and the foundation of its progress. Consequently, despite the importance allotted to the agrarian reform, it is undoubtedly true that since 1940, industrialization has constituted the main concern of economic policy. And until very recently, it has been unanimously supported by economists. There is no opposition, only a divergence of opinion regarding the pattern of the process and the sectors which should be given priority. Moreover, in the latter part of the 1950s, and even today, once the industrial base had been established, one topic has been subject to frequent discussion—foreign investment, which has experienced radical change in structure. Indirect investment has increased while direct investment has changed from being a supplier of raw materials and public utilities to being concerned with manufacturing for the growing Mexican market.

A second topic of importance that began to attract the interest of many economists in the sixties has been that of protectionism for the promotion of industry.

The interest was born from the participation of Mexico in the Latin American Free Trade Association and the plans for economic integration. The efficiency of the Mexican industrial apparatus, as well as its competitive position in world markets, became a matter of concern. Another factor, to be discussed later, which also contributed to this, was the stagnation of exports of primary products toward the end of the 1950s. Since then, there has been a great preoccupation with increasing exports of manufactured goods in order to raise the rate of growth of exports, in view of the limited prospects for agricultural exports.

Foreign Investment

The traditional type of foreign investment, oriented toward the exploitation of natural resources and public utilities, which had become firmly entrenched in Mexico by the turn of the century, practically disappeared after the oil expropriation (1938).

In recent years, direct foreign investment has established itself basically in the manufacturing and service industries, oriented toward the domestic market, abetted by the policy of import substitution and high profits. This has also been the result of the Mexican economic policy of reserving certain basic sectors of the national economy for the State and leaving others to the private sector. Moreover, there is a requirement of a majority of Mexican capital in those sectors considered as intimately connected with basic industries.

With regard to the composition of direct foreign investment, despite the recent diversification of the countries of origin, that of the United States accounts for 70 percent of the total. As far as the balance of payments is concerned, there has been a trend toward a growth of remittances abroad in relation to the influx of new capital. The observation, resulting from a comparison of a stock of capital with part of the income derived from its use, is supplemented by another element—the payments to foreign capital for such items as interest, royalties, and technical assistance grow at a faster pace than the profits. Within the Mexican economy, an increase has been experienced also in the reinvestment of profits of foreign companies in relation to new foreign investment.

Finally, since 1942, indirect foreign investment, in the form of international loans for development, has become increasingly important, especially for the financing of social overhead capital projects executed by the public sector.

The opinions of some Mexican economists with regard to the results of foreign investment and the measures they recommend for its treatment are presented below. However, it should be remembered that the appraisal of foreign investment in Mexico is often associated with ideological positions or the defense of vested interests. One reason lies in the lack of objective studies on the subject. The most serious and impartial works are either descriptive or make a very general analysis.

The first opinions to be presented are those of the experts who are in favor of the broadest possible scope for foreign investment, followed by those which express the viewpoints of the entrepreneurial and banking interest. A work by G. R. Velasco (1955) illustrates this current of thought. It states that foreign investment is beneficial, for it introduces new technology and creates new sources of employment for local labor. The author believes that it has been insufficient mainly for two reasons. First of all, there are fewer sources of capital today than before the war (1955). And secondly, there is a diffidence about investing in the Latin American countries because of their instability. Because of the requirements of economic growth and the

scarcity of capital, Mexico needs foreign investment—it is the only thing that can raise productivity and accelerate development. For that reason, Velasco recommends an "open-door" policy. And, upholding a liberal economic concept patterned after the nineteenth century model, he considers direct foreign investment as the most convenient, for it guarantees the freedom of private enterprise and development. On the other hand, indirect foreign investment is unstable, has a political character, increases the influence of government in the economy, and leads to a less efficient use of resources (p. 18).

A diametrically opposite viewpoint is held by the Marxist left. A work by J. L. Ceceña (1965) can be considered representative of this current of thought. His approach is two-fold, as he repudiates two forces, international imperialism from without and the ruling oligarchy within a capitalist, underdeveloped economy. Foreign investment is the bond that ties and maintains the power of both of these oppressive groups. Those who uphold this position also manifest a very intense nationalism, and, on occasion, seem to be unaware of the structural change experienced in the pattern of foreign investment in Mexico that has taken place over the past three decades (F. Carmona, 1963). The main arguments are that foreign investment is an obstacle to development because it makes the economy tend toward disequilibrium and make it vulnerable to changes abroad. Moreover, it is conducive to a dangerous dependence on the United States, both in the markets for Mexican goods and as a source of supplies for national industry.

The nationalist intent makes this group of economists argue that economic independence is identified with industrial power, and foreign investment is harmful because it introduces disadvantages to the industrialization process. For one thing, the acquisition of modern techniques of production and business administration is limited, for the foreign subsidiaries operate with secondhand processes, which reduce industrial efficiency. It also absorbs smaller national firms or subjects them to the market conditions established by foreign firms. Furthermore, it gives place to a marked economic concentration which permits the use of monopolistic practices that only bring about domestic price increases. It is stated that of the 400 largest firms operating in Mexico, foreign enterprises make 36 percent of total sales. With supplementary data, it can be deduced that there is a "dual" economy, in which more than half of the firms are operated by head offices abroad. Thus, economic planning and industrial development are impeded, for the long-sought integration of national industry is only achieved by the head offices and their subsidiaries.

The leftist economists also find that banking makes for foreign dependence for reasons tied to foreign investment. Commercial banking is dominated by strong foreign interests and is greatly concentrated. Centering around the most important banks, a number of enterprises have been established that supplement, in financial terms, a complete system of economic power. Thus, banking is an instrument used by foreign monopolies to make use of domestic savings. This, in turn, has three consequences: easy access to the liquid resources of the country, a greater influence on business, and ties with the interests of national capitalists (Ceceña, 1965, p. 292). Moreover, the effect on the balance of payments represents a disinvestment for the country, because of excessive outflow of profits in the form of foreign exchange. A number of corrective measures have been proposed. To begin with, the leftist economists assert that economic policy must have two well-defined goals. On an international

level, the country should pursue an independent policy trying to establish closer relations with Latin America. On the national level, the policy to be followed should be revolutionary and nationalistic, and should lend support to Mexican entrepreneurs by means of credit, technical aid, etc. Consequential to their belief in state intervention in all of the economy, however, they demand a policy applicable to foreign investment, with special norms, since at present the interests of the country are not defended. The great monopolies must be combatted by means of legislation and a planned regulation.

The third viewpoint, eclectic and more pragmatic, is held by a group of economists who, except for transitory differences, can be considered to follow the present official position. In general, they admit that foreign investment has complemented domestic savings, for it has contributed to doubling the per capita gross national income since 1939.

If the official statements on the subject of foreign investment were analyzed, it would be found that there are definite policy norms. Foreign investment must not displace national capital, which will have exclusivity in basic activities and majority control in those branches connected with the basic industries. It must be subject to the laws of the country. The preferred priority is: inter–government loans, international loans and direct foreign investment. Of the last, those that would be best received are the ones bringing technological innovations.

According to Alfredo Navarrete, who is an exponent of this position, there are other supplementary criteria. Direct foreign investment, which is now concentrated in the manufacturing industry, is geared to supplying the local market. This change in its orientation has made for a greater rigidity in the balance of payments, because formerly, with the traditional pattern of foreign investment, the remittances of earnings moved parallel to export income (Navarrete, 1966). The association of national and foreign capitals is an efficient procedure to attain technological progress. Local investors participate actively in the determination of profitable fields, while foreign capital has the technical responsibility. Mexican economic policy is in favor of the fusion of capitals with the object of increasing domestic capital formation. Actually, this has been the result not only of policy but also of the economic maturity of the country, due to the existence of an entrepreneurial as well as an investing class that favors this process (Navarrete, 1966). In the financial sense, foreign investment has not altered the debtor status of Mexico, although it has a smaller debtor position today than in 1939, because the increase in the foreign debt has been invested in productive works that enhance the payments capacity of the nation (Navarrete, 1958). The inflow of capitals is advantageous to both the recipient and the investing country. The effects of development loans on the balance of payments depends on the productivity of the investment, the rate of interest due on the loan and the length of time for which the loan is extended. In the case of direct foreign investments, these effects depend on the rate of profits and the reinvestment policies of the investors. The recommendations of Navarrete for the appraisal of foreign investment are based on the following criteria: a) complementary (in terms of domestic saving); b) flexibility (in the composition of investments); and c) equilibrium (between domestic and foreign saving and between direct and indirect investment). He points out that Mexico needs capital and for that reason it should also take advantage of the institutional saving of the insurance companies of the developed countries as another means of increasing the

resources available for investment. But it is preferable that public investment be financed with domestic saving, generated by the voluntary savings of firms and individuals (Navarrete, 1958).

It would be interesting to mention two additional points regarding foreign investment. The first of these is the difference in the standpoints of the two private industrial sectors, the Confederation of Industrial Chambers (CONCAMIN), which represents big business, and the National Chamber of Manufacturing Industry (CNIT), representing medium and small industrial firms. The second point is related to an essay by M. S. Wionczek (1968) regarding technology, foreign investment and derived problems.

The CONCAMIN standpoint, as expressed by Campillo (1966), is fairly close to the official position. Foreign investment is a useful supplement to domestic saving in fostering development. It should not displace national capital and, preferably, a partnership of the two should be formed. It should be oriented to activities of social interest, providing jobs for Mexican technicians and personnel, as well as introducing new techniques. To judge it, it is necessary not only to consider reinvestment and the remittance of earnings but also to take into account the substitution for imports, the increase in exports and the multiplier effect of (foreign) investment on the economy in general (Campillo, 1966, p. 13). He does not believe that it would be beneficial to have a law to regulate foreign investment, but he does establish the need for a foreign investment policy that would unify the criteria of the various government agencies.

The appraisal made by the CNIT is just the opposite. Its exponent, R. A. Ollervides, asserts that domestic saving has been the real lever for development, not foreign investment (1966, p. 491). The latter could serve as a support if there were legislation to regulate it. Domestic saving is not insufficient, as bankers claim, but it is poorly channeled because of the excessive amounts of credit granted to commerce to the detriment of industrial activities. It is the banking system which is insufficient. There is no need to divert attention to other areas while defending vested interests. Finally, he argues, foreign investment has accentuated the exchange disequilibria and disinvestment of the country. Referring to a statement of the bankers, seconded by CONCAMIN, he maintains that it is difficult to assess the beneficial effects of foreign investment on the current account of the balance of payments. Moreover, it "generates additional imports that not only mean additional drawings of foreign exchange . . . but through clever devices used to import foreign machinery, equipment and material . . . deprive domestic activities of demand necessary for their development" (Ollervides, 1966, p. 490). The same author adds that the effect of foreign investment upon the ". . . so–called current account of the balance of payments is indirect and its influence on it very difficult to measure" (p. 490). International loans have other disadvantages, such as tied credits, which represent a subsidy for the creditor nation and impede industrialization on a national scale (Ollervides, 1966). For this reason he demands state intervention, so that, on the one hand, the basic and related industries be nationalized or Mexican–owned, and on the other hand, foreign investment not be completely abolished, but strictly regulated by means of appropriate legislation, since the present standards are not adequate for the interests of the country.

How can the difference in attitudes of the private industrial sectors be explained? Some grounds could be found in the conflict between manufacturers and

importers that existed toward the end of World War II. Nevertheless, the present friction between small and medium producer and big business and banking is the outgrowth of the process of industrialization itself. Two disadvantages for the former that do not exist for the latter can be deduced from the statements of Ollervides—the scarcity of credit and the difficult access to external sources of intermediate goods. In substance, they seem to feel that industrial development is leaving them behind, and at the mercy of the big firms, because of their position not only in the market for goods, but also in that for factors of production. For this reason, they turn to the protection of the State, asking for its intervention from a standpoint reminiscent of that of the Marxist left, wherein the defense of interests and nationalist elements are combined with market arguments.

A second issue of interest is that of the technological gap and foreign investment. Wionczek (1967) begins by pointing out that within Mexican industry, foreign investment has shifted toward the most profitable and dynamic sectors. This has increased its profits and reinvestment, which could jeopardize the balance of payments position were the process to be reversed (no new investment and the repatriation of profits).

The increasing dependence of the country on foreign technology implies a new cause for friction, in view of the inequality of strength of large corporations and Mexican entrepreneurs (see the position of the CNIT). This is disturbing because "second-hand technology is received at an exaggerated price" and because technological dependence is associated with the introduction of new techniques by means of private foreign investment (Wionczek, 1967). Yet, due to "the quasi monopolistic conditions of the firms, it would be naive to ask for the liberalization of the conditions of the transference of new technology and its divorce from the investment of "capital" (p. 985).

Industrial Protectionism

The process of industrialization has been closely linked to protectionism since its beginnings in the nineteenth century. It was not until the Second World War, however, that protectionism became the basic instrument of industrialization. The larger demand during the war, resulting from the increase in exports and the budget deficit caused by the increased spending on public works, could not be satisfied entirely, regardless of the much greater utilization of the installed capacity of firms. Thus, the inflationary process became more marked. Moreover, there was some negligence regarding the standards of quality of manufactured goods.

Toward the end of 1945, the Law for the Promotion of Manufacturing Industry (Ley de Fomento de Industrias de Transformación) was enacted. It defined new and necessary industries and provided a stimulus for newly-formed basic manufacturing enterprises. The balance of payments problems, resulting from the disequilibrium of prices at home with those abroad, as well as from the deferred war-time demand, made it necessary in 1947 to establish a system of import permits, especially for luxury goods. R. Izquierdo (1964) states that at the end of the 1940s there was great enthusiasm for protection. The government championed industrialization and appeared to have found the formula for sustained and accelerated growth. Moreover, requests for protection from foreign competition were willingly considered, without careful attention to the type of product or the extent to which national materials were employed in its manufacture (p. 268).

The decade of the 1950s was characterized by two different phases of protection-

ism, separated by the devaluation of 1954 and the price stability attained subsequently (1957). After the devaluation of the peso in 1954, two goals of economic policy became apparent: to avoid domestic price increases and to provide incentives for foreign investment. The latter was justified in that protectionism had already altered the structure of imports, leading to the predominance of imports of capital goods. This, in turn, represented a greater inelasticity of demand for imports, which made manifest the dependence of industry on supplies from abroad. After the attainment of price stability, protectionist policy, based on the instrument of import permits, came to have as its fundamental objective domestic industrial integration. Import substitution and the consumption of domestically-produced goods was stimulated thereby. The system of licensing imports proved to be very effective because of its flexibility.

Opinion regarding the protectionist policy is a consensus that it has been necessary and beneficial for the goal of industrial development. However, its usefulness as an isolated tool is coming to an end and it should be reformulated in view of the present conditions of development of the country. Conditions are no longer such that by merely limiting foreign competition there can be a promotion of economic growth and industrial diversification. In the future, substitution for imports will be effective only in conjunction with the integration of national industry (Izquierdo, 1964).

In general, also according to Izquierdo, the Mexican experience shows that the formula of industrialization through protection is not self-perpetuating. Furthermore, the government has almost always given favorable replies to requests for protection, it has done so without due consideration to the type of product or its proportion of imported inputs, and without demanding the fulfillment of progressive integration programs. What might be called the "natural" theory of import replacement was widely accepted (Izquierdo, 1964, chapter 4). On the other hand, P. Garcia Reynoso stresses that industralization was attained at the expense of other sectors, e.g., agriculture, as well as through the sacrifice of the consumer. In relation to its effects, Garcia Reynoso also mentions some distortions wrought by protectionism on the Mexican economy. He specifically points out three outstanding problems: A) High costs, resulting from an indiscriminate protectionism, have rendered manufacturing industry inefficient. This particular trait limits demand in the domestic market and reduces the competitive position in foreign markets (1968, p. 963). (J. Saenz [1957] had already pointed this out in 1957, when refuting the ECLA theory on external disequilibrium, to which we will refer below). B) Geographic concentration of industrial activities. C) The limited impact of industrialization on total employment. Saenz criticizes the idea that industrialization per se overcomes unemployment, because it overlooks demographic growth and the intensity of the technological revolution. He indicates that this problem has become more acute lately because the technology that Mexico absorbs comes from highly developed countries where labor is the scarce factor and uses capital-and technology-intensive methods of production. Therefore, as long as Mexico improves her industrial structure and develops basic and intermediate industries, the marginal increments in employment are shrinking relative to the rapid growth of the labor force (Garcia Reynoso, 1962, p. 964).

The industrial development policy has not been free of problems. Izquierdo indicates that in terms of the process followed in its implementation, there have emerged

two important problems. First of all the special problem involved in shifting the demand for foreign final goods to a demand for intermediate goods was not taken into account. Secondly, the lag between the emergence of a potential market for import substitutes and the effective decision of the private sector to satisfy that market was ignored (Izquierdo, 1964). In addition the divergent goals of the public and private sectors—and of the pressure groups within each sector—added other complex factors that obstructed the rational practice of protectionism.

Furthermore, Izquierdo notes that the government has been concerned by the pressures on the balance of payments resulting from the cumulative increase in costs and prices wrought by the restrictive measures, which put them above those prevailing on world markets. The alternative, he maintains, is to act decisively on private investment. The restrictive measures continue to be a regulating factor, for their indiscriminate use would cut off the supply of necessary goods for industry with subsequent damage to domestic demand (Izquierdo, 1964).

There is agreement on the need to revise protectionist policy. For Izquierdo, the subordination of an import policy to an over-all plan for investment in industry cannot be delayed. This need inevitably arises when instruments of protection operate independently, unrelated to any long-range objectives (Izquierdo, 1964, chap. 4, p. 283). A long-term plan would divert the evaluation of the two types of restrictive measures—permits and tariffs—from the examination of their immediate repercussions. At the same time, a closer coordination between the Secretariat of Industry and Commerce and the Secretariat of Finance is required; these secretariats administer import permits and tariffs. The measures proposed by García Reynoso also show a general feeling that has emerged recently. In a broad sense, they are oriented to the solution of the distortions he analyzes, based on the continued promotion of industry. More specifically, he calls for the revision of the protectionist mechanism and points out that the system of permits should be subject to considerations about the level of prices, the use of domestic inputs, and the effect on the balance of payments (1968).

Some interesting conclusions can be inferred from the works on protectionism. In the first place, it is accepted, without exception, that the protectionist policy has been indispensable for the growth of industry. While its shortcomings are readily admitted, this does not invalidate the fact that it has been the main instrument for the establishment of the industrial apparatus. In other words, the issue was whether or not to create an industry, and this issue was very important for nationalist economic policy. In the second place, in the late 1950s, when the rate of import substitution began to decline, there existed a basic industrial framework, the country entered a period of exchange stability, and pressures on prices were lessened; a new subject of increasing interest arose—industrial efficiency, measured in terms of costs in relation to world prices. Last but not least, there is no talk of eliminating protectionism in order to attain efficiency and confront the domestic producer with world competition. Instead, it is considered a matter of perfecting the method of licensing and selecting imports in accordance with the broader scope of industrial integration, linked to an investment policy.

It is worth noting that no analysis has been made of the role of direct foreign investment in the process of import substitution and, consequently, how foreign firms have benefited from the protectionist trade policy—a benefit that goes against the nationalist interests of the pol-

icy's makers. Lastly, the idea that industry has still not matured fully is maintained implicitly. It is noted that Latin American economic integration can provide stimuli on the market side and could be a means of selecting the most efficient firms.

IV. *The Distribution of Income*

The distribution of income shifted in a regressive way during World War II, because of the diminishing relative share of wages and salaries. The possible causes of redistribution, different in their nature and effects, that attracted the attention of Mexican economists above all were agrarian reform and inflation. The interest in the latter quickly surpassed that in the former. In the 1950s—a period of rising prices—numerous and interesting works appeared on this subject, including an essay by I. M. de Navarrete, in 1960, which is the best known work on the distribution of income in Mexico. It is unanimously agreed that there is a very inequitable distribution of income, even more so than formerly. Different observers present different reasons for this, but some points in common can be found. Initially, the following section will present different viewpoints on the inequitable distribution of income. The causal factors will be examined next. Finally, some of the measures proposed to correct the inequitable distribution will be considered.

In her study, Mrs. Navarrete asserts that the distribution of income was more inequitable in 1960 than it had been twenty-five years before. That is to say, the salaried and farming sectors have experienced a decrease in their relative share within national income. This assertion is supported by the conclusions of a mission from the World Bank which established the same point for the period 1939–1950 (Birf y Nacional Financiera, 1953). The same writer also indicates that there is a process of concentration which only benefits capital owners. These opinions are shared by almost all of the economists who have dealt with the subject, such as V. Urquidi, Navarrete, J. Noyola, Flores de la Peña, Carmona, B. Siegel and others.

Because of the inflationary method followed by the government, development policy tends to accentuate the inequitable distribution of income. Such is the opinion of various authors, including Mrs. Navarrete (1955). Moreover, these authors believe that this is inevitable, for in an initial stage of development it is necessary for economic policy to foster the formation of capital. Only at a later stage will the efforts directed at enlarging the domestic market solve the inequitability. As it will be shown later, many feel that Mexico has already reached this stage. In short, the consensus of opinion is that there has been a process of concentration of income in favor of profit earners. This concept, which was dominant during the inflationary period, has not been refuted in the recent years of exchange stability.

Several explanations are offered for this type of income distribution. Actually, the various opinions supplement each other and center on economic and institutional factors. The first group to be presented here is that concerned with the effects of the economic variables involved in the production function.

In the early 1950s, two complementary articles appeared in *Trimestre Económico*, which approach the problem from the standpoint of salaries. One was a joint study by Noyola and D. López Rosado (1951), and the second was by Flores de la Peña and A. Ferrer (1951). The former two pointed out that the level of real salaries had deteriorated, despite the increase in the real per capita income of the country and in productive efficiency in every activity, but that the income derived from

this increase had gone to the holders of capital. The determining factor had been a surplus in the supply of labor (excess supply). The other two authors mentioned above note that the owners of capital have experienced additional gains through inflation, since they kept the profits generated by the increased productivity. Urquidi (1961) of the structuralist school, derives the inequitable income distribution from the characteristics and effects of the supply conditions of some factors of production, such as the rigidity of agricultural output, the insufficient rate of industrialization, and the high rate of growth of the supply of labor. Finally, Mrs. Navarrete (1959) also attributes the disparity of incomes to low productivity, which can be explained by the fact that within the occupational structure, there is a predominance of agriculture, handicrafts and small proprietors.

Included among the economic explanations are inflation, the characteristics of the tax system, and the situation in world markets.

According to the consensus, the inflationary process had a redistributive effect in favor of the owners of capital, although no empirical proof is presented to substantiate the statement. What is really being pointed out is that the rise in prices accentuates the inequitable distribution or constitutes an obstacle to corrective measures. No one contends that without inflation there would be an equitable distribution. The characteristics of the tax system are singled out as another causal factor, taking into account that it is regressive. The tax burden is not shared equitably among taxpayers. The predominance of taxes on expenditures and the absence of taxes on investment of limited social value prevent the tax system from acting as a mechanism of redistribution of income. Many experts agree on this, for example, Mrs. Navarrete, Urquidi, Flores de la Peña and others. They all attribute the situation to the fact that the objective of equitable taxation has not been applied properly by the fiscal authorities, which have been mainly concerned with tax collection. There are also those who link the problems of income distribution to the conditions of world markets, like Flores de la Peña who maintains that in export activities the transfer of income from the labor sector to the capital sector is promoted sometimes by currency devaluations; another factor which also has an adverse effect is the price increases in the world markets. (Flores de la Peña & Ferrer, 1951, p. 621). Last of all are the so-called institutional factors. These can be summarized fundamentally by the lack of organization among workers or their limited bargaining power. Moreover, after 1940, wage policy ceased to be as aggressive as it had been formerly, which also helps to explain the low level of real wages and the advantageous position of capital. On this topic the absence of disagreement among Mexican economists is evident. They agree on the skewness of income distribution, and their opinions about the causes of the unequal distribution are complementary. There is one question remaining: how to mitigate the inequality.

Naturally, most of the authors argue for a better distribution of income and advocate action on what have been termed causal factors. In their attempt to achieve this objective, they want a preponderant role for government measures. The intervention of the state for the promotion of economic development, applicable to the productive apparatus, would abolish unemployment. The concomitant increase in productivity, together with monetary stability, would be to the advantage of real wages. Such is the opinion of Navarrete (1955) and others who are concerned about economic growth. The preferred instrument, undoubtedly, is fiscal policy

(despite the fact that it has been shown that it is very limited as a mechanism of redistribution in industrialized countries). Mrs. Navarrete considers that because of the disparities generated by inflation, fiscal policy should absorb a substantial part of profits and channel them toward the kind of investment that raises productivity (1955, p. 245). In the early 1950s, a radical reform of the tax system was sought, although now the emphasis is on improving and perfecting the system. In 1951, Flores de la Peña and Ferrer (1951) were clamoring for the elimination of the regressive nature of taxation by means of an increase in direct taxes. Six years later, Urquidi [1956] urged the same thing. The former of these authors also recommended the establishment of additional taxes on exports to absorb the exchange profits of exporters, which increase considerably in the event of a devaluation. As a direct measure to increase the income of wage earners, they proposed a more aggressive wage policy and a greater mobility of the wage scale. Similarly, references are made to the objective of improving the bargaining position of wage earners. These opinions are expressed by Mrs. Navarrete and Flores de la Peña, but they do not propose concrete measures for the removal of the institutional factors that promote an inequitable income distribution.

Most of the proposed measures would either speed up or delay economic growth, depending on the point of view one has of economic development. In other words, whether the stimulus for investment must come from saving, fostered by raising the level of profits, or from demand, fostered by enlarging the market to guarantee the absorption of the goods produced and based on mass consumption. For some, then, the inequitable distribution of income guarantees the process of capital formation. Others, however, closer to Keynesian ideas, believe that the development of Mexico is now limited by the insufficiency of the domestic market. A better distribution of the national income, besides meeting the requirements of social justice, is warranted by the objective of accelerating the process of growth.

The argument for the enlargement of the domestic market rests on the inequitable distribution of income and on the limited demand for consumer goods industries, some of which do not utilize their full capacity. The first of these reasons has to do with distribution, while the second is related to production. The evaluation inferred from this situation will depend on the predominant goals of economic policy. Proof of this can be found in recent statements of the Minister of Finance, A. Ortiz Mena, who asserts that fiscal policy must be defined in terms of the formation of the domestic market, without which economic development cannot acquire sustained momentum. A healthy and fair development, he maintains, requires that the inequalities of income be corrected (1966, p. 9).

Among Mexican economists, this standpoint is not new. Combined with many elements of Keynesian policy, it already had a following during the inflationary period, which was characterized as stemming from an excess demand. For example, *Revista de Economía*, representative of current economic thought, in its editorial of May, 1954, pointed out that the crucial problems of domestic economic development have as their basic reason the constant contraction of the domestic market because of the smaller relative incomes of the groups with a high propensity to consume. If that is so, the effects of the peso devaluation (1954) and future policies must be evaluated and conceived on the basis of this conditioning factor. If the premises are valid, the future policy of the Government must reenforce as much as possible the purchasing power of the

masses of the national population, not only for reasons of social justice and equality, but as a central foundation for a theory of economic growth. Urquidi agrees with this idea and also mentions the insufficiency of the domestic market as a result of the structure of the distribution of income. The concentration of income, due to inflation and structural elements, results in a structure of total demand which does not seem congruent with the possibilities of full employment of the capital of the country (Urquidi, 1956, p. 435). There, then, is one reason to enlarge the domestic market. Summing up, this idea has been shared by a great number of economists. It has also been the basis for the explanation of inflation expounded by the structuralists, as will be shown later. There is a divergence of opinion on the same subject. For instance, R. Hoyo D'Addona (1968) states that in Mexico there is still an excessive potential demand in terms of total supply, and that therefore, the fiscal system should stress the capitalization of the country.

Another phenomenon that has been observed within the structure of demand has been the change experienced as a result of increasing urbanization. According to Fernandez Hurtado (1960), the exodus from the rural areas to the urban centers was due to the agrarian reform. Yet, whether it was prompted by the armed struggle or the distribution of land, it had no economic justification such as would have been provided by an increase in the productivity of land and a subsequent excess of agricultural output relative to rural consumption. The increase in the urban population has marked a new direction in the structure of the demand for the goods of domestic industry and has generated an extraordinary boom in the urban construction industry, a net consumer of national goods (Fernandez Hurtado, 1960, p. 202).

V. *Stabilization Policies*

Monetary Aspects

Like other variables of the system, the monetary behavior of Mexico since World War II can be divided into two distinct phases. The first, until 1957, was marked by inflation; the second, from 1957 to date, has been characterized by stability. An inflationary process began in the Mexican economy in 1935, during which time money grew proportionately to income. From that time until 1956, the average annual increase in prices was 11 percent, although with wide fluctuations, especially during the war and the periods immediately following devaluations. However, in the period between 1956 and 1968, price increases have decreased asymptotically and have become moderate. Similarly, changes in the money supply are now more than proportionally related to the growth of money income.

In the beginning, the inflation had mixed characteristics of cost increases and increases in demand. Later on, the latter was the predominant characteristic, for the public sector obtained funds by inflationary means and promoted inflation, as its expenditures constituted an excess demand in relation to productive capacity.

Foreign trade was another factor in the process. Exports of agricultural commodities—the only significant exports—grew very rapidly due to price increases and, especially after the agrarian reform and the agricultural development works that followed, also by production increments generating thereby rises in the money supply. Nevertheless, in the period of stability which began in 1957, and was associated with smaller increases of agricultural output, there was a decrease in the rate of growth of these exports. The income-elasticity of the demand for imports also fell from one period to the other, but the policy of protectionism

changed the structure of imports so that those of capital goods and intermediate goods became predominant.

The devaluation of 1938 was related to the expropriation of oil; that of 1948–1949 was considered to be part of the worldwide parity changes that took place in the postwar years. On the other hand, the devaluation of 1954 was an isolated event which had a drastic effect on national expectations. Once more it was promoted by public spending, executed by a new administration which, using Keynesian criteria, tried to counteract the decrease in business activity (attributed to the recession in the American economy in 1953 after the Korean War) by means of deficit spending financed by the central bank. The devaluation, which was described as a preventive measure, established the exchange rate at its present value of 12.50 pesos to the dollar. Conditions had changed as compared with the preceding devaluations. The system had become more flexible in terms of output especially in the agricultural sector. This, together with the more extensive use of foreign loans for public investment after the devaluation, meant that the central bank no longer had to finance the investment of the public sector. Thus the depreciation of 1954 was a transitory inflationary influence, since the economic variables that affected the level of prices followed a trend toward equilibrium and determined conditions of growth with price stability. Later on, beginning in 1958, saving in financial institutions tended to grow more rapidly, as price expectations changed, so that it was possible to finance the new government deficits with private savings, placing government securities in the banking system to supplement foreign loans. In this process, the greater use of reserve requirements was very important. This instrument was also applied to non-monetary intermediaries, which were the institutions that registered the fastest growth. It also made it easier for the Banco de Mexico to attract genuine savings to finance public investment. To recapitulate, the former monetary expansions were replaced by the use of domestic and foreign savings. The earlier public investment contributed to higher output, and there was also a change in the structure of consumption and an increase in the propensity to save. In effect, the initial elimination of the government deficit, and the different way of financing it employed later, abated the inflationary demand pressures stemming from public finance.

The Inflationary Period

The inflationary process has produced a diversity of opinions and controversies about monetary policy. In the first part of this section, attention will be given to the interpretations of the monetary phenomena that occurred during the inflationary period, followed by the different explanations of the causes of the inflationary situation itself. In order to facilitate the presentation, events will be grouped together in two chronological periods: those which center on the devaluation of 1948–1949 and those which took place around the time of the devaluation of 1954. The dividing line has been established as the Korean War, in 1950, for it put an end to the effects of the first devaluation and began to generate the causes of the second.

For a majority of the experts, the inflationary process immediately following World War II was the result of two factors. The first of these was the strong drive toward public investment with deficit financing. The second factor was the growing deficit in the balance of payments, attributed to the excess of imports that corresponded to the deferred demand of the war years.[5]

[5] For some authors, however, the second factor was of an external nature—the end of World War II.

Were there other conditions that led to the devaluation of 1948–1949? A classic liberal viewpoint was expounded by a banker, A. de Iturbide (quoted by E. Padilla, 1948, p. 397), who pointed out that State intervention was a cause of the disequilibrium for it brought about reforms that resulted only in price increases and decreased agricultural output. Among other factors were: the agrarian reform and financing *ejidatarios* and small proprietors; the aid given to production cooperatives; and the oil expropriation and the nationalization of the railroads. Another viewpoint was based on the application of the purchasing power parity theory to the relative levels of prices in the United States and Mexico. Saenz is representative of this current of thought, and he adds to this analysis the influence of the stock of money. Refuting an argument of the Economic Commission for Latin America (ECLA), he stated that the peso had been overvalued since 1941, and that this fostered imports and the balance of payments disequilibrium. If Mexico did not devalue before (1948) it was for reasons derived from the war: a) the control of production in other countries; b) the amount of capital which found a haven in Mexico and meant an abundance of dollars despite the exchange disequilibrium, and c) the remittances of dollars by Mexican farm labor in the United States, the sum of which was as important as exports (E. Padilla, 1948, p. 405).

The devaluation did not occur until 1948 because of the Government's fear of the consequences. It did not want to devalue before industrialists could purchase abroad at low prices, which they had not been able to do during the war. Another reason was the case of obtaining foreign loans, and still another was the hope that a point of parity would be reached autonomously as a result of the inflation in the United States (Sáenz, quoted in E. Padilla, 1948). According to Gómez, Director of the Banco de México (1964), the overvaluation of the peso was brought about because prices rose more rapidly than in the United States between 1945 and 1948, and when price controls were lifted in that country, it was impossible to correct the disequilibrium in the balance of payments by reducing domestic demand. He added that the greatest effort was made to maintain the exchange rate by using reserves and foreign credit. But the offers of further credits required a balanced budget, which was impossible in the short run. For this reason, there finally was a devaluation (Gomez, 1964, pp. 778–779). In an attempt to place the blame for the devaluation, Sáenz holds all of the economic groups responsible. The blame was shared by the government with its equivocal monetary policy which made production costlier; the private sectors because of their policies of hoarding and of speculation, and especially the industrialists, for not improving technology and relying on the support of excessive protectionism; labor for demanding high salaries; and private banking, because of the voracity of some bankers. In short, the situation brought about a rigid cost structure, which could not be reduced, far above that which should have prevailed with an exchange rate of 4.85 (E. Padilla, 1948, pp. 406–407). Finally, he adds, monetary policy between 1945 and 1948 was equivocal, in allowing the dwindling of reserves in the interest of maintaining the exchange rate and because the new exchange rate only left the peso with a very slight undervaluation, which was quickly neutralized. What Saenz therefore recommended was a more austere monetary and credit policy, to be pursued within more restricted margins, since the currency was chronically overvalued with respect to the rate of exchange. This made it necessary to continuously limit the quantity of money in

order to avoid serious repercussions on the reserves of foreign exchange (Saenz, 1957). It would also be necessary to revise protectionist policy, for excessive protection distorts the struture of domestic costs and thereby fosters overvaluation and price instability. Therefore, this policy, including tariffs and subsidies, permits and quotas, should be administered much more carefully than monetary and credit policy (Saenz, 1957, p. 538).

Another current of economic opinion is to be found in the opposition from the left. Often they introduce external factors as causes for the disequilibria and are strong partisans of exchange controls. According to E. Lobato Lopez (1968), the orientation of Mexican economic policy has always been predominantly Keynesian and monetary policy was actually a policy of forced saving, which concentrated collective saving into a few hands, in order to encourage production at the expense of consumption. N. Bassols (quoted in E. Padilla, 1948) focused on political and external factors. He maintained that if the inflation was not constrained at the right time, it was for political reasons and because all hope converged on foreign aid–loans or foreign investment. He felt that this policy should be abolished and than attention should be paid to the country and her internal solutions. The same line of reasoning is followed by R. Torres Gaitán who emphasizes external factors, and more specifically the attitude of the United States. He notes that American purchases during the war brought about an increase in the money supply which led to inflation. Later on, when the United States began to sell to Mexico, problems emerged in the form of balance of payments deficits. Therefore, Mexico should have revaluated her currency during the war, and once the war was over, devaluated it (quoted in E. Padilla, 1948, p. 402). Noyola (1948) made the observation that the rise in prices stemmed from the demand side as well as the supply side. On the demand side, the two outstanding factors were the growth of the money supply and deferred demand. The pressures generated from the supply side pertained to the increases in the prices of imported goods, due to their scarcity because of restrictions on imports; the rise of prices in the United States, and the effects of the devaluation of 1948.

These economists view the devaluation in a different light. According to Bassols, the monetary reserve was not used in the best possible way, that is, to purchase production goods, instead of which much of it was squandered (quoted in E. Padilla, 1948). Torres Gaitán distinguishes between external and internal causes of the devaluation: the external factors are based on the decadence of capitalism and the policies of the United States and the IMF. Those of an internal nature include the balance of payments disequilibrium, the price differentials between the United States and Mexico, the flight of capital, the deferred demand for imports, and the complete failure of import controls. Of all of these, the disequilibrium of the balance of payments was the most important. In any case, he concludes, the devaluation was extemporaneous (in E. Padilla, 1948). Finally, A. Noriega Herrera (1955), indicates that the short–term results of the devaluation were positive, but he casts a doubt on the timeliness of the measure, asking whether it would not have been preferable to establish exchange controls or whether the monetary authorities carried out the devaluation, believing in the "optimistic promises of the International Monetary Fund." The negative results readily appeared, since the price of imports increased in domestic currency, commodities became more expensive for the domestic consumer be-

cause of the preference to sell export goods abroad, and inflationary pressures emerged in all economic activities (Noriega, 1955, p. 159).

The solutions proposed by Torres Gaitán can be considered representative of this current of thought. He maintains that if the country's exports could be increased, and if imports could be reduced, not only the balance of payments would be righted, but also domestic activity would be stimulated as a consequence of more benefits stemming from the increase in exports and the decrease of goods that can be done without (in E. Padilla, 1948, p. 401). He subsequently proposed price controls, exchange controls, and tailoring foreign trade policy to agricultural and industrial policies. Noyola (1948) was more inclined to find the primordial causes of the price increases in the real factors of the economy. He stated that a policy of price controls would be tenable only if there were controls on the volume of commodities, and not the "police" measures that had been used until now. Moreover, it would be convenient to apply, in a supplementary policy of moving wage scales linked to the cost-of-living index. Nevertheless, success would depend on monetary and fiscal policy and an uncontrollable factor–American inflation (1948, pp. 9, 10).

Until now, the opinions presented have concerned the events prior to the Korean conflict. Opinions on the events that led to the devaluation of 1954 will be examined next, and then the end of the inflationary process and the stability attained between 1957 and 1958 will be reviewed.

Gómez, an exponent of the official view, interprets the devaluation of 1954 in the following way. The excessive credit generated by the banking system to finance the public sector deficit, as a result of following Keynesian-like advice to offset the effects of the 1953 American recession, drained the foreign exchange holdings of the central bank. But on this occasion the policy pursued was different from that of 1948—the peso was devalued before the monetary reserves were depleted and before resorting to greater indebtedness to the IMF and the U. S. Treasury (1964, p. 780). On the other hand, W. Sedwitz (1957) stresses phenomena in the external sector that preceded and provoked the devaluation. What precipitated the devaluation of 1954, he claims, was not the absolute level of the foreign exchange reserves, but the rate at which they were dwindling (1957, p. 12). The deterioration was due to a) the decline in the volume and the price of exports, as well as to b) the slowdown in business activity in the United States. These were the two most important factors. c) Moreover, the deterioration was accompanied by an increase in imports as a result of the credit and fiscal expansion which constituted an attempt to offset the effects of the American recession; and d) the situation was worsened by the outflow of short-term capital, made possible by internal liquidity and the fear of devaluation or exchange controls. The outflow of capital in 1954 was promoted by the excessive liquidity of the monetary system and by the ease of selling government securities without capital losses and of placing profits into dollar accounts in Mexican or American banks (1957, p. 23). This is why a restrictive-depressive policy was difficult to maintain and perhaps the monetary authorities themselves wanted to count on wider margins to carry out the anti-cyclical and development policies. According to Sedwitz, it would be unfair to infer that the devaluation was equivocal, since the positive phenomena that followed had still not taken place and because the pressures on the dwindling reserves of foreign exchange and domestic depression were very strong.

The leftist current also made new contributions. It is possible that it was strongly influenced by the publications of the ECLA during that period, which had not existed at the time of the previous devaluation. The reference here is to concepts that later gained wide acceptance, such as the theory of the deterioration of the terms of trade of Latin America and the dualistic approach of R. Prebisch (ECLA, 1950), who divides countries into two large groups—the cyclical centers and the peripheral nations. These economists continued to stress the external sector, the idea of a cyclical center providing an explanation for the origin of all the cyclical disturbances in the peripheral countries. The views expressed by Noriega (1955) are an example of this approach. To him, the Korean War was the reason for the disruption of the relative stability that had been attained, for it brought about a series of phenomena that altered the entire Mexican economic system. Among the most important changes were the inflow of foreign capital, the rise in the world prices of raw materials, and considerable increases in receipts of some invisible items of the balance of payments, that cooperated to trigger domestic inflationary pressures. Among internal factors he stresses the lack of skills of those in charge of economic policy and the absence of a measure to limit the remittance of profits by subsidiaries of foreign companies to their countries of origin. The measures chosen to abate the inflation were not successful because of the new trends in the countries that are cyclical centers, such as rearmament, the speculative boom, the contraction in the raw materials market, and the increased tax burden (Noriega, 1955). Concerning the balance of payments disequilibrium, he pointed out that it was a result of the worsening of the terms of trade, since export prices fell 14 percent while the prices of imports rose 0.8 percent (1955, p. 163). The solutions offered do not differ from those that had been proposed for the devaluation of 1948. Thus, this writer stated that devaluations have proved incapable of correcting the periodic disequilibria of the balance of payments, much less of offsetting them. The solution lies in multiple exchange controls, which he considers advantageous because they limit the profits of exporters and adjust imports to the available foreign exchange. They also protect the cost of living, permit the increase of the revenues of the State, and encourage tourism and new exports (1955, p. 168).

Taking a position contrary to official policy (and within the field of the structuralist current) Lobato (1968) stated that inflation had two fundamental effects —price increases and a reduction in the purchasing power of the majority of the population. Economic policy is not effective for it cannot reduce prices or even curb price increases, nor can it augment the standard of living of the population. Since taking action on prices is difficult, the solution would have to be the redistribution of income, by means of a decisive policy of raising wages and salaries. Technically, he says, this policy is valid, for it would enlarge the domestic market and foster development.

The inflation gave rise to different viewpoints on its causes as well as on appropriate solutions. Opinions were very definite, frequently based on the writings of a favorite author. There was one current of opinion, the structuralist, opposed to public policy, which formulated a novel explanation of the inflation and which when its proponents were joined by other Latin American economists came to represent a new "school" of thought. Others remained within the bounds of an orthodox interpretation, even though they disagreed on the importance of the determining variables.

The Structuralist Position: The most outstanding spokesman of structuralist thought was Noyola, who was the first to systematize all the pertinent tools of analysis in 1956. The current was developed in conjunction with the staff of economists of the ECLA, in Santiago de Chile. The work of M. Kalecki at the United Nations Secretariat was influential upon the economists that rallied to structuralism. The controversy between ECLA and the IMF about inflation in Latin America influenced the way of thinking of many.

Noyola readily admits the influence of two European economists, Kalecki and H. Aujac. The former underlines the importance of the rigidity of supply and the extent of monopoly power in the economic system. In contrast, Aujac focuses on the behavior of the various social classes and their bargaining power. Noyola summarizes his position briefly: inflation is not a monetary phenomenon; it is the result of disequilibria of a very real nature that are expressed in the form of increases in the general level of prices (1956, p. 604). He introduces a series of factors to examine inflation in Latin America and divides them into three groups: those of a structural nature (population and productivity); those of a dynamic nature (different rates of growth in the diverse sectors of production); and those of an institutional character (the behavior of the public and the private sectors).

Noyola points out that among the factors which produce inflation there are two distinct kinds. One of them is made up of what he calls the "basic inflationary pressures," on the magnitude of which depends the degree of inflation. These pressures are due to disequilibria of growth, which are almost always to be found in the slow growth of two sectors, foreign trade and agriculture. The second group consists of "mechanisms of propagation" which have secondary or spread effects, for they only heighten the inflationary process through the performance of the economic system itself. The mechanisms of propagation can be of three kinds—the fiscal mechanism, the credit mechanism, and the mechanism for the readjustment of prices and incomes. It can be said, then, that the inflationary process is born from the emergence of the basic inflationary pressures that result from the disequilibria arising from the efforts of growth in the economic system. Once the inflationary conditions are established, inflation is further stimulated by the mechanisms of propagation. Some conclusions inferred from the model are applicable to the Mexican reality. Since 1935, the basic inflationary pressures in Mexico have stemmed from the inability of exports to grow at the same speed as the domestic economy. This has led to disequilibria in the balance of payments and to the subsequent devaluations (Noyola, 1956, p. 611). Yet, the increasing diversification of exports has abated the force of these pressures. With regard to agriculture, Noyola inferred something else: that there was almost no inflationary pressure stemming from the rigidity of agricultural supply, that is, the supply of foodstuffs (1956, p. 612). This was the result of the agrarian reform and the agricultural development policy. However, in the short run, agriculture had played a role in that sense, for example, when the crops had been poor (Noyola, 1956).

The distribution of income and the behavior of social groups are manifest in the mechanisms of propagation. Noyola remarks that in the case of Mexico these factors have been weak, and that the weakness has been even more pronounced among the lower-income groups. Taxation had tended to become more regressive because of the increases in prices, and the role of current expenditures as a redistributor of national income has been insuffi-

cient. The credit mechanism had been the most passive in the propagation of inflation, which only went to prove the well-known axiom that monetary policy is only effective when it abates business activity, increases unemployment or curbs development.

The mechanism for the readjustment of prices and incomes was the most important of the three in propagating inflation, because of the high degree of monopoly that exists in the Mexican economy. With the exception of rental housing properties, in almost all other prices of goods and services, entrepreneurs were able to transfer the incidence of inflationary pressures, although salaried groups could not do the same (Noyola, 1956, p. 614). This proved the weakness in the adjustment of salaries, which stems from structural and institutional reasons. The structural reason was to be found in the existence of a surplus army of agricultural labor with a very low productivity, which tended to depress the level of real wages and to debilitate labor organizations (Noyola, 1956, p. 615). The institutional reason lay in the relationship between organized labor and the public sector, that is, government paternalism that weakened the labor movement. To recapitulate, Noyola identified two factors as the principal causes of inflation in Mexico—the foreign trade disequilibrium due to insufficient exports as the basic inflationary pressure, and the degree of monopoly as the mechanism of propagation. His conclusion was that if a choice must be made between inflation and economic stagnation or unemployment, inflation is preferable. The negative side of inflation is not the increase in prices, but its effects on the distribution of income and the distortions that it wreaks in the productive structure and the structure of demand. It may be impossible to restrain inflationary pressures, but they can be abated by means of a very progressive fiscal policy, as well as price contols and wage adjustments (Noyola, 1956, p. 616).

Another author, Flores de la Peña (1953), introduced a few other elements in his analysis of the Mexican inflation. He differed from Noyola in that he considered the inelasticity of agricultural production as one of the determining factors of instability. He also identified, as an inflationary pressure, the increase in public investment, which he considered to have had the opposite effect than Noyola propounded, i.e., it fostered the expansion of output. Flores de la Peña differed again on the issue of the role of exports, for he felt that the growth of exports constituted another source of inflationary pressures, because it brought about the boom in foreign investment, a greater sensitivity to international economic contractions, and an inadequate specialization in basic commodities. According to Flores de la Peña, the growth of exports brought about an increase in the quantity of money, and therefore, had an inflationary effect. Noyola, however, considered that it was the insufficiency of exports vis-a-vis the increase in imports that created the balance of payments disequilibria that led to the devaluations and the changes in the level of prices. Within Noyola's reasoning, there is evidence of the theory of the deterioration of the exchange rate by external reasons, such as the decline in the prices of exports. Flores de la Peña proved that the effect of the balance of payments on inflationary pressures had not been significant. He maintained that the influence of the increase in prices abroad or the devaluations on the general price level had been overstated. He felt that it ought to be admitted that the inflationary pressures had been caused by gross investment, or, especially, by its composition (1953, p. 470). It is precisely on investment that he based his explanation of the inflationary process. Public investment created inflationary

pressures because of the lag between its monetary effect and its real effect. Private investment creates these pressures when it is "poorly oriented," that is, when it is directed toward superfluous goods or for foreign markets. In general, "the expansive capacity of investment will be greater as the proportion of income generated passing to the lower-income groups is greater" (1953, p. 467). Besides, the origin of inflationary pressures lies in the low productivity of private and public investment. The inelasticity of supply plays a role in the mechanics of inflation, according to Flores de la Peña. The inelasticity of supply is due to the inelasticity of domestic output and the scarcity of foreign exchange. The reasons for the former have already been examined in the section on agriculture. The scarcity of foreign exchange is a phenomenon that is not uncommon in a period when there are substantial imports of capital goods, which prevents covering the deficiency of output with commodity imports. Consequently, the inflationary process due to economic development begins when an increase in the income generated by gross investment raises effective demand to the extent that it brings about a scarcity of factors of production. In an underdeveloped country, this point is easily reached because of the inelasticity of supply (1953, p. 477). Thus, inflation can break out before the economy reaches full employment, due to the fact that any increase of income gives rise to a more than proportional increase in the demand for foodstuffs and manufactured goods for mass consumption, and rarely does the growth of investment mean a significant expansion of the output of these goods (1953, p. 462).

Until now only the differences of opinion of the two authors have been discussed, but they share some opinions. Their similarity lies in that they both examine inflation in its real form. They reject the monetary approach and monetary policies and they imply that inflation is inevitable in the process of development. Both have been influenced by Kalecki, and both share a concern about the effects of inflation on the distribution of income. In other words, both authors study inflation by comparing total effective demand with total supply, stressing the real aspect. They consider that monetary policies have not been effective because they could not prevent the increase in the money supply. According to Flores de la Peña, monetarist policy seeks a balanced budget and prefers monetary and credit policies, to foster economic development. But it must be understood that the quantity of money is merely a result and that by itself it does not determine anything. The solution is simple: reduce demand or increase supply. Thus, monetary policy only affects demand and as a means of reaching equilibrium is an absurdity in a country that has a very limited domestic demand (1953, pp. 472–473). These ideas supplement those of Noyola, who states that monetary policy is effective only when it strangles economic development (1956, p. 614). The solutions given by Flores de la Peña are directed basically to acting upon investment. Thus, the intervention of the State acquires a decisive function, for it must establish priorities for investments, trying to maintain the equilibrium between effective demand and total supply. Moreover, for the reestablishment of equilibrium, he advises finishing the agrarian reform in order to improve agricultural production and eliminating the insufficiency in the supply of foodstuffs. He suggests that raising wages and salaries is useless because prices would increase at a faster rate, and that the only way of maintaining the purchasing power of the income of the urban population is by increasing the supply of foodstuffs and manufactured goods for

mass consumption. If the process of the concentration of income is allowed to go too far, a point would be reached where any reduction of the real wage would be accompanied by a more than proportional decrease in the effective demand for manufactured goods. He asserts that this has happened in Mexico in basic industries, like the textiles and shoe industries. Flores de la Peña does not see any advantage in price controls, because decrees do not increase output, and it is precisely upon output that economic policy must have an effect (1953, pp. 479–480).

Although it is difficult to really place him within any current of thought, Torres Gaitán (1962), presents a viewpoint considerably different from those of the other two authors. Acknowledging that the external disequilibrium is the result of a domestic demand which exceeds domestic supply, he feels that the solution lies in increasing the latter as well as raising national productivity. The insufficiency of domestic supply helps him to explain both the deterioration of the exchange rate and the basic inflationary pressures, in the same way as the structuralist approach. Torres Gaitán adds that the adjustment has been made by means of different instruments. Before World War II, the main instrument was beggar-my-neighbor policies. More recently, the most important instruments have been controls on foreign trade and the use of foreign loans. He points out specifically, moreover, that the permanent disequilibrium of the exchange rate is due to exports of basic commodities with a negative income elasticity of demand made more acute by the encroachment of synthetic substitutes, and by the excess supply of primary commodities that depresses their prices (1962).

In its analysis of external equilibrium in Latin America and regional foreign trade problems, the ECLA studied the case of Mexico, with the collaboration of Noyola, C. Furtado, O. Sunkel and Urquidi. In synthesis, they asserted that the devaluations had occurred not because the exchange rate was overvalued, but for structural reasons. Structural reasons predominate as a result of economic development itself, that is, the incongruence between the pace of increase of total supply and total demand. The disequilibrium is even more acute when development is accompanied by a concentration of income, due to the high income–elasticity of imports of durable goods (United Nations Economic and Social Council, 1957).

In practice, they present the import function as dependent on the rate of growth of income, which tends to produce foreign disequilibrium when there is an acceleration of income growth; disequilibrium is accentuated by the deterioration in the terms of trade (United Nations . . . , 1957). Devaluations do not alter this combination of causal factors, because it is a structural phenomenon of the economy, related to the inelasticity of supply (especially that of agriculture), which plays a role in making the response to increases in external demand quite slow. Nevertheless, the higher foreign exchange income is reflected in higher domestic spending, price increases and increases in imports, until the value of exports falls because of a change in the terms of trade, and the current account deficit is heightened. Thus, if devaluations were effective in the past, for example, in stimulating exports of certain goods, such as cotton, or in fostering the substitution of imports of other goods, later on the range of these possibilities became much narrower (United Nations, 1957).

The next part of the analysis was the object of controversy. In order to prove that the external disequilibrium had structural causes, it was deduced, from a series of price indexes based on the purchasing power parity theory, that the exchange

rate had been undervalued since 1937. In other words, the external disequilibrium could not be attributed to an overvalued exchange rate.

The Orthodox Position: Sáenz (1957) disagreed with the thesis of ECLA. He considered that the overvaluation of the rate of exchange during that period had been the central problem. Besides pointing out that ECLA should have defined what was meant by an equilibrium exchange rate, he noted that between 1937 and 1955 the balance of payments had undergone periods of equilibrium and dynamic stability, which did not appear in the tables elaborated by that organization. On this basis, Sáenz maintained that it was wrong to speak of a period of overvaluation or undervaluation between 1937 and 1955. Moreover, the ECLA thesis has some technical shortcomings. Within the theory of purchasing power parity, it is misleading to use a very long period of time because of the variables that are omitted from the estimation, which with time may become more important than those that have been included (1957). Similarly, he asserted, ECLA used a base year which was not representative. If the periods had been divided into more logical stages, the conclusions would be different. Besides, greater refinements should have been used; for example, sectorial and special price indices, changes in the degree of tariff protection or variations in the elasticities of exports and imports (1957). In conclusion, the data, presented in a more orderly fashion, would indicate that the peso was overvalued in the period under consideration and for that reason it has been more predisposed to shocks or disturbances. For Sáenz, the increase in prices did not stem from the absence of a balanced budget, except to a small extent, for the expenditures had been held at prudent levels. The real causal factors were the generation of income from development itself, the expansion of bank and extra bank credit, and the general increase in costs due to commercial policy. The first and the last of these were inevitable (Sáenz, 1958).

One group of economists considered that external factors are fundamental in explaining monetary disequilibria. Some of them shared the thesis of ECLA, but without accepting the production rigidities and while considering the influence of monetary variables. The argument was that public expenditures, or a surplus on the current account of the balance of payments, produces an increase in aggregate demand, and induces increases in investment and income, and that imports rise in an even greater proportion than income. Since the growth of exports is slow—and they might even decline because of the deterioration of the terms of trade—a deficit on current account appears. Change in relative prices—domestic and foreign—or a monetary policy oriented toward sustaining output growth, accentuates the balance of payments deficit and it becomes necessary to devalue to restore external equilibrium. This approach utilizes the type of perverse elasticity of demand for imports that was so fashionable among economists after the end of World War II. (Siegel, 1960, and J. S. de Beers, 1953, estimated an income-elasticity of imports of close to 2, from which they inferred that it was necessary to export more, use more foreign loans or devalue periodically. This viewpoint was shared by Navarrete.)

In his analysis of the Mexican inflation, Siegel used orthodox instruments to determine what kind of inflation there had been. Although he is not a Mexican economist, his work, like that of de Beers, influenced the discussion of Mexican economists. This author studied the goods market and the demand for money in order to locate the source of inflationary pressures, both domestic and external. He came to

the conclusion that inflation in Mexico was not a cost inflation, and that the Mexican experience stands as proof that industrialization is not necessarily accompanied by a lag in agricultural output, as some structuralists maintain, for agricultural production grew more rapidly than industrial output and population (Siegel, 1960). It can be said that the periods with the largest price increases coincided with the years of the devaluations. In turn, it would seem that the only real pressure on the cost side related to external factors, apart from devaluations, came through raw materials and production goods and was associated with the strong rise of import prices after the war. Pressure on the wage side was insignificant, due to the weak bargaining position of unions. Hence Siegel concluded that the Mexican inflation was not a cost inflation, and that if it was so in any way, it was because of the increase in import prices and not because of differences in the rates of growth of the various sectors (1960, p. 81). There is no doubt that price increases were the result of the effects of demand. The causes of variations in the money supply, Siegel said, indicated the origin of inflationary pressure (1960, p. 82) and can be identified as internal and external. In general terms, it seems that between 1941 and 1948, internal factors were mainly responsible.

The internal factors most important in generating inflationary pressures were the government deficits and the loans granted by state-owned banks. The deficits were a direct result of the inability of the government to finance its investment programs by means of taxation or non-bank loans (Siegel, 1960, p. 88). It is also true that inflation was tied to development because of the maintenance of a very high rate of investment. Although they were not in themselves inflationary forces, there were institutional and cultural factors that accelerated the inflationary process, for example, the level of technology or the structure of consumption. Furthermore, the weakness of the instruments of monetary regulation was evident, the reason for this lying in the limited development of the money markets, in a banking system that was not very independent, and in inflationary fiscal policies. Siegal adds that monetary regulation is not effective when the government cannot practice budgetary discipline while instituting a development program. In general, the monetary authorities have been more tolerant of inflation than of deflation.

However, provisionally, and at times decisively, the effect of external factors was felt intensely. Such was the case in the periods 1942–1945, 1950–1951 and 1954–1955.

In the periods of devaluation, the pressure coming from external factors was more important. The devaluations of 1948–1949 and 1954 were preceded by declines in international reserves that resulted from excessive imports (Siegel, 1960, p. 93). Siegel is also of the opinion that between 1951 and 1953 the deterioration of the real terms of trade brought on new problems. In short, the external sector has played a very complex role. On the one hand, it has brought about a growth of income through exports. On the other hand, through imports, it has led to balance of payments deficits.

Regarding the solution of the problem of external disequilibrium, the ECLA group believed that it was highly improbable that new devaluations would have positive effects to counteract the contraction of demand or income. An acceptable policy would be the reorientation of productive resources based on the development of demand and on the capacity to import. This could permit gradual structural changes in supply, directing investment and fostering the required substitution of

imports (United Nations, 1957). Devaluation, a measure that has been used to counteract external disequilibrium, is a *dangerous instrument for the correction of* the balance of payments in the short run. ECLA was also of the opinion that devaluation could not correct the fundamental causes of the disequilibrium. Its immediate and long-run effects could accentuate the inflationary process. Other factors, such as the high rate of domestic spending, the tendency toward budgetary deficit, and the weakness of the instruments of monetary control in the face of international disturbances can cancel out the initial advantages of devaluation.

In general, it can be said that inflation in Mexico did not become a hyperinflation, for the characteristic mechanisms acting to increase velocity did not appear. Nor was there a price–wage spiral of any significant proportion. It was, perhaps, conditions in the labor market itself that prevented hyperinflation.

The Period of Stability

After 1955, there was a decline in the number of works on monetary policy. Those published toward the end of the decade, when stability was well entrenched, still referred to the recent inflationary problem. In effect, as the influence of inflation waned, so did the number and intensity of the writings of economists about monetary stability. This waning continued through the 1960s, and the focus of economic discussion moved to other areas.

Sedwitz (1957) attributes the recovery that began after the devaluation, in the second half of 1954, fundamentally to the expansion of exports, but argues that, contrary to what might have been expected, the devaluation had a very small effect on this turn of events. What elements, then, came into play? The same author points out two. One, external in nature, was the recovery of business activity in the United States, which even resulted in an increase in dollar prices in the world market. The other, of a domestic character, was the extraordinary crops resulting from optimal climatic conditions, especially those of cotton and coffee. The phenomenon can be explained by the fact that there is a high price–elasticity of demand for exports. However, exports are more sensitive to income changes abroad than to prices, for they consist mainly of raw materials. The devaluation had little effect on exports because it did not change their dollar price, or the volume of exports (1957, pp. 14–16). The influence of the devaluation in reducing imports was not important, owing to the low (less than unitary) price–elasticity of the domestic demand and the high elasticity of external supply. There was a substitution effect with regard to durable consumer goods, the demand for which was channeled to national industry. Imports of raw materials increased because of the stimulus provided by the devaluation to local industry. Actually, the change in the composition of imports was due above all to the increase in agricultural output, because of the reduction in imports of foodstuffs, as well as to the greater restrictions on imports (Sedwitz, 1957). This analysis is diametrically opposed to the hypothesis of the inelasticity of agricultural supply of Flores de la Peña.

Some experts arrived at the conclusion that the devaluation did not play a decisive role in the recovery of the economy after 1954. The basic reasons for the recovery were the good crops and the boom of the United States economy. Secondly, there was an increase in tourism, on which the devaluation of 1954 had a positive effect.

Many authors abstained from passing judgment on the prospects for stability, remaining on the level of offering solutions and confining themselves to the short run.

Others foresaw the continuation of the inflation, because the basic causes of the problem were not being checked. However, Sedwitz, in 1957, also noted signs that would guarantee future monetary stability. His reasoning was that economic policy has three fundamental functions related to monetary stability: 1) to prevent the outflow of capital; 2) to maintain the stability of the exchange rate, and 3) to make anti-inflationary policy effective by means of a control on the volume of money and prices. In terms of each of these functions, there was a favorable outlook for the role of economic policy, both from the domestic and the foreign points of view (1957).

According to Gómez, the period of adjustment of the money supply and prices lasted until 1960. In the years of world prosperity between 1955 and 1957, the Banco de Mexico recovered its reserves, capital returned to the country, and there was an increase in foreign and domestic private investment, because of the confidence inspired by free convertibility. With the growth of domestic saving, fostered by stability, a big step was taken in the financing of development. Similarly, public investment in social overhead capital was supported by all of the concurrent events (1964). Transitory difficulties could be solved satisfactorily by resorting to the monetary reserve as an offsetting factor. More recently the increase in bank financing of the private sector has reinforced the downward trend of the rate of interest. Interest rates had risen drastically in real terms when price increases ceased, even though there was not much change in the nominal rates. Gómez adds that in the struggle against inflation, the objective was never stability per se (1964, p. 778). In a developing country, inflationary pressures are harder to overcome because there is a smaller capacity for voluntary saving and for the payment of taxes, at the same time as financing requirements are larger (1964, p. 780). Later on, Gómez mentions an institutional factor influencing the performance of Mexican economic policy. For the monetary authorities, the most difficult pressures to resist are those coming from the public sector. Government officials want to advance as much as possible while they are in power, at the expense of other sectors of the administration (1964, p. 78). This is how budget deficits and inflationary pressures are created. Moreover, state-owned enterprises might be another source of pressure, because of negligence in matters of efficiency and productivity. The pressures caused by the private sector originate mainly in the banking system, because of the banks' interest in obtaining reduced reserve requirements as well as unlimited rediscounts. Yet, the central bank maintains that what is lacking is voluntary savings and is opposed to unnecessary raising of the levels of financing. Other pressures coming from the private sector include requests for the elimination of taxes and the granting of subsidies (Gómez, 1964). Economic stability fosters a higher volume of investment and exchange stability is an indispensable element in economic progress.

Ortiz Mena, Finance Minister, also mentions the achievements attained through stability. Saving has experienced a marked growth, for it has been duly channeled by banking institutions toward the kind of investment required by development. Institutional saving was not the only savings source to grow, for the securities market has expanded also. There was an increase in the volume of securities and with it, a slight decrease in the rate of interest. The government considers that domestic savings are the best means of accelerating development (Ortiz Mena, 1963). The objectives of monetary and financial policy, then, are to enhance the

mobility of domestic resources and prevent any inflationary tendency.

VI. *Fiscal Policy*

In the post-war period, fiscal policy has been characterized by the predominance of public spending, in view of the goals of development, as well as by traditionalist features of taxation maintained until the present decade. Public spending increased because of the pressure of developmental requirements that called for sustained public investment. Adhering to its doctrine of the State as a promoter of development, the government undertook the social overhead capital projects that were characteristic of the postwar period.

Indirect taxes still play an important role in the structure of taxation, giving it a regressive bias. The tax adjustments of 1955, 1962 and 1965 led to a lesser dependence on taxes on foreign trade; indirect taxes were reduced while direct taxes were increased, and an attempt was made to foster the formation of capital. Actually the income tax reforms did not alter the general structure of incentives for the capitalization of firms. Three distinct periods, more or less characterized by tax reforms, will be examined here in order to present the fiscal policy of the postwar years. The first of these periods covers from 1945 until the reforms of 1955; the next, up to the reforms of 1962, and the last period, from the 1965 reforms up to date.

Tax Changes Up to 1955

In examining the tax reforms of 1955, and after having suggested some guidelines for future investment, Urquidi (1956) attached great importance to the facility of capitalizing reserves, which fosters investment by allowing a firm to increase its capital. Capitalizing reserves also allows banks to turn to other credit requirements, discouraging the distribution of excessive profits, and consequently, conspicuous consumption. Other aspects of the reforms of 1955 dealt with depreciation allowances, which facilitated the rapid amortization of equipment. The same author also considers that the deductions permitted for donations and cultural appropriations lend support to education and research. Nevertheless, Mrs. Navarerete (1957) expressed the belief that the formation of capital attained by the private sector could have been achieved to the same extent without a preferential system of taxation, except for tariffs. She arrived at this conclusion because of the characteristics of the tax system itself, and the inflationary process. Among other things, she mentioned that the level of taxes was very low, that there were numerous ways of evading taxes, that taxes were paid only on a part of income, and that the entrepreneur could easily pay taxes because of inflated profits (1957, p. 38). Urquidi also has doubts about the system of exemptions to new and necessary industries, for they are not significant in the determination of industrial location and they represent a loss of revenue for the public sector and benefit only a small group of firms.

Regarding the effect of fiscal policy on demand, the criticism is more severe. According to Mrs. Navarrete, the objective of a more equitable tax burden had been reached only in a very limited way: since efforts had been dedicated to the collection of funds, they had not embraced the highest goal of Mexican policy, economic development with social justice (1957, p. 38). She also came to the conclusion that the present system of taxation was regressive. In agreement with this criticism, Urquidi added that the income tax did not fall on the combined income of the individual and that it should be reformed, for it perpetuated an undesirable distribution of income and fostered the production and importing of luxury consumer goods (1956).

Both economists recommended a

broader, more progressive tax system to promote the financing of public investment. Mrs. Navarrete demanded a redistribution of the tax burden in accordance with the economic capacity of the tax payers. She also recommended that voluntary private savings be stimulated by direct financial measures and not indirectly through the absence of a personal income tax (1957, pp. 41–42). Urquidi also insisted on the need for greater tax revenues in order to meet the requirements of public investment and welfare services, such as schools, public health and so forth. According to him, the economic development of Mexico should rest, to a great extent, on an adequate and equitable taxation of income.[6] He specifically underlines the desirable characteristics of the income tax: that it is equitable, that its effects on investment and consumption should contribute to the promotion of development, and that it represent a significant and stable source of funds for the State.

The above suggests that there are no dissident opinions. Public spending, with its goal of promoting the development of the country, is viewed favorably. (Except for the extreme liberal standpoint, which does not, which does not tolerate state intervention.) Evaluations of the tax system generally agreed, that it was inadequate for the needs of the country: it was regressive, insufficient, and had operational shortcomings. Authors like Flores de la Peña and Noyola also felt that it should stimulate a greater domestic demand. These were the first manifestations of the two currents of thought that are in disagreement about the strategy of development. One maintains that action must be taken to encourage savings and the formation of capital, while the other argues that capital formation is the result of a strong domestic market. Both points of view agree on the need to reform the tax system, even if the proposed measures are different.

The Period of Tax Change of 1962–1964

Throughout 1959–1961 the Mexican economy experienced a setback, in the form of a decrease in the rate of growth of the national product due to the stagnation of private investment. This happened during the first years of attainment of price stability. The government increased its level of investment in face of the alternative of a real recession. Deficit financing, with which government spending was still operating, can be attributed to the shortcomings of the tax system, which made the problems of financing and the smallness of the securities market more acute. In order to overcome the problem, the government made revisions—not a radical change to eliminate obsolete tax laws—having as its objective to make the tax burden more equitable by gathering some sources of income into more general basic groups. It also tried to build in more incentives for private investment. Among the main tax changes were the introduction of accelerated depreciation for new investments, the widening of the tax base and an increase in the progessivity of tax rates. The economists' comments on the changes were complimentary rather than antagonistic. Although some did so with reservations, the critics justified the reform because it was a step towards changing from the former cedular system to a global system of taxation on some incomes, with cumulative rates for incomes coming from different sources, while maintaining exemptions or special treatment for certain incomes from capital. The reservations stemmed from the reforms being viewed as insufficient to meet the needs of the country.

[6] Elsewhere, he added that the apparatus for revenue collection was slow and lacked coordination. And since the tax burden is not equitable—for its incidence is greatest on those engaged in productive activities—he points out a contradiction in the tax system. On the one hand, private investment is encouraged, and on the other, it is discouraged through taxation.

Concerning personal income taxes on income from remuneration of personal services or returns to capital, Mrs. Navarrete felt that many shortcomings were left, despite the reforms. There were loopholes in the way interest and returns on financial securities, as well as rental income from real estate and capital gains, were taxed. There is no system of deductions which would alleviate the tax burden on lower-income groups. The progressive aspect is very limited because there is no rate for total personal income. There was still great inequality and a good deal of discrimination in the distribution of the tax burden.

A. Cervantes Delgado (1962), in particular, maintained that there were shortcomings in the structure of the income tax, such as the breakdown of income through the cedular system of sources of income, and there was still no equitability in taxation. There was no application of the basic principle that all of the income of people must be taxed regardless of its source (1962, p. 402). According to Cervantes, the trend toward the reduction of indirect taxes was more illusory than real, due to the shortcomings of the cedular system. He asserted that the system of taxation lacked flexibility; thus, it was shown as unable to stabilize economic fluctuations in the short run. Increases in the volume of revenues had been due to administrative reasons, such as a change of tariffs coinciding with the exchange effects of the devaluations, the enforcement of new tax laws or the granting of facilities to taxpayers in arrears (1962).

Among others, Cervantes, and C. Ocaña and J. Alejo (1939a&b), were of the opinion that Mexico is one of the countries where the rate of taxation is lowest. They added that because of the undesirable distribution of income there was no reason for maintaining such a small tax burden. Yet, this situation could not be remedied by improving the cedular system; there was no alternative except to institute a single tax on income. Some of Kaldor's comments on the subject (1963) gained wide acceptance. Kaldor, who was later a consultant to the fiscal authorities in the implementation of the reforms of 1965, felt that the fact that only a small part of the gross national product was captured by taxation was due, not to poverty, but to the extent of inequitability in the distribution of income (the high concentration of wealth), as well as to the regressive nature of indirect taxes. All of this, in summary, was a reflection of the inability to tax effectively the most prosperous sectors of the community. Furthermore, in practice, many direct taxes were not enforced —which can be seen by comparing the data on national income with income declared for tax purposes. There were also ample legal means for tax evasion through exemptions and omissions. A case at hand is that of the anomymous ownership of securities (more often than not these are made out to the bearer), which is an obstacle to the collection of taxes on property.

In general terms, the criticisms were aimed at the fact that the owners of capital remained in a veritably privileged tax position. Yet taxes on labor income meant that employees and workers bore the largest tax burden, besides which the collection of these taxes is the most effective because tax withholding is practiced. These ideas were held by, among others, Ocaña and Alejo who added that the basic breakdown of incomes prevented the application of higher tax rates; that even though the rates were higher now, a redistribution of income had not been possible, and that the effects on wages and salaries had hampered the development of a domestic market (1963b, p. 173). Mrs. Navarrete also argued that the tax bias in favor of the owners of capital, besides its inequitable

nature, had a disadvantageous effect on the objective of development, because the owners of capital are stimulated to invest in securities with fixed returns and low taxes, instead of taking risks in new enterprises.

The Tax Reforms of 1965

The reforms of 1962 were considered insufficient. They were the object of much criticism and there were many recommendations that influenced the changes that took place in 1965.

According to Mrs. Navarrete, there are three relevant basic principles: that all incomes should be taxed, regardless of their source; that labor's ability to pay taxes is less than that of capital, so that a system of tax deductions should be worked out in order to tax in accordance with the ability to pay; and that a policy should be formulated to lend progressiveness to the tax system. Cervantes also indicated that the fiscal system should be used to supplement private saving, and that taxes should be applied with the objective of stimulating or discouraging given activities, according to their social value (1962, p. 406).

Among others, Mrs. Navarrete (1964) and Ocaña and Alejo (1963a) stressed once again the need to make fiscal policy fulfill three objectives—stimulate production activities (supply), improve the distribution of income (demand), and collect revenues without jeopardizing monetary stability or the balance of payments. Mrs. Navarrete advocated a favorable treatment for those who carry out productive activities or risk capital (1964).[7] C. Ocaña and J. Alejo (59) added that it is necessary to avoid the transference of taxes by shifting from the high-income groups to the low-income groups, since the former both demand and supply goods and services, while the latter demand consumer goods and supply personal services. Thus, Mrs. Navarrete was concerned with the promotion of output, and the other two concentrated on the distribution of income.

Among other things, Kaldor suggested that it was necessary to effect fiscal reforms for two reasons—the insufficient level of public revenues and the fact that the inequitable distribution of income could lead to political problems. He recommended that the family be considered as the basic economic unit for tax purposes. He also suggested that corporate profits be taxed at a single rate of 40 percent and, as a general criterion, that a mixed system be applied that taxed income as well as wealth, pointing out that he would consider this more just and equitable because it approached the capacity for payment of the individual. Hoyo D'Addona of the Secretariat of Finance disagreed with Kaldor on the form that direct corporate taxation should take. He felt that it would be better to maintain the principle of progressiveness. The justification for his argument was based on the fact that the structure of corporations in Mexico is such that a few people could amass large profits because of the limited distribution of shares among the public. Hoyo D'Addona added that there was no pressing need for an overall personal income tax, although he admitted the necessity of changing the cedular system of taxation (1968).

With respect to income taxation, it was said time and again that it should be applied to total income. The authors mentioned above clamored for the consolidation of the cedular system: the tax to be applied to personal income would act as a single rate, based on a concept of income

[7] In other words, she did not believe it expedient to collect high taxes from firms, which are the organizations that carry out productive activities and investment; what should be taxed is the income of entrepreneurs.

that would cover all of the earnings of the individual. Kaldor, for example, called for two changes in the rates of taxation: higher levels of tax–exempt incomes and higher tax rates that would be more progressive (1964, p. 25). The same idea had been propounded by Ocaña and Alejo in 1963 (b).

Kaldor's preference for a tax on both income and wealth was shared by Mrs. Navarrete. A tax exclusively on income does not reflect the economic capacity of the individual, which also depends on his property assets (1964). According to Kaldor, the tax on the profits of a firm would be a means of taxing the income of its owners, and he recommended a tax exemption on undistributed profits. The essential concept of income as discussed by him includes personal income as well as corporate income. As far as personal income is concerned, after insisting that there was a bias in favor of the owners of capital, Kaldor suggested that the tax on income from wages and salaries be made annual and that generated by capital be modified in order to tax capital gains. Regarding corporate taxes, he felt that all capital gains should be taxed, that deductions should be permitted for losses (which could be reabsorbed by later profits) and for depreciation (wear and tear), that tax exemptions for "new and necessary" industries be abolished, and that, instead, direct subsidies be granted for fixed capital expenditures. Among other changes recommended by Kaldor were to abolish the distinction between non–commercial and commercial transactions; to do away with the anonymous character of the ownership of bonds, shares and obligations, and to introduce a tax on wealth, together with an overall tax on income (1964, p. 33).

Some authors regarded the policy on the administration of state property as a supplement to fiscal policy, a source of revenue and an instrument of development. Cervantes (1962), demanded that the state administer the country's resources in a more economical and uniform way, be they untransferable assets in use, or assets in the form of investment assets.

The reform of 1965 brought about an innovation that had repeatedly been requested—the transformation of the cedular system of taxation into an overall income tax. Nevertheless, opinions about fiscal policy have not changed very much. The problems are alleged to remain the same, in general terms. Those engaged in the study of these problems are concerned mainly with the formation of capital and the redistribution of income. Let us consider some comments of Hoyo D'Addona and Ortiz Mena made in 1968 and 1966, respectively.

Hoyo D'Addona, who examined the effects of fiscal policy on supply, argues that in order to promote economic development, fiscal policy should stress the use of public expenditures. Understating the time–lag between investment and its effects, especially because of the long gestation periods of the social overhead capital projects executed by public investment, he asserts that public spending is not inflationary because it is offset by the production of goods and services. Similarly, according to the same author, the concept of deficit can vary, depending on whether there is a preponderance of investment expenditures or of current spending on government consumption. He implies that the former are advantageous for economic development, although he prefers that public spending be financed by means of taxes. All of this is in agreement with a previous statement, in which he claimed that in Mexico, there is still an excess of potential demand in relation to total supply (1966). Therefore, it develops that this author emphasizes the idea that fiscal policy should seek, above

all, the formation of capital in the country. He provides an argument for encouraging the reinvestment of profits by means of a tax exemption for undistributed profits. For the same reason, he advocates the simplification of the tax schedule for the small taxpayer—a tax preference in favor of the small and medium sized firm—in order to attain a harmonious and balanced development.

The Secretary of Finance, Ortiz Mena, holds a similar view: fiscal policy should reenforce saving and investment. The most suitable instruments of that policy—in terms of private investment—are the tax system and the strengthening of the securities market. Similar goals should be pursued by tariff policy. Reflecting the recommendation of Kaldor among others, Ortiz Mena maintains that the structure of the tax on the income of firms should be so designed that it fosters the financing of capital formation by such devices as exemptions for undistributed profits and accelerated depreciation. With regard to the effects on domestic demand, Ortiz Mena points out some aspects that already have been mentioned, for example, that the redistribution of income and the enlargement of the domestic market are the established objectives for fiscal policy (1966). He notes that the progressiveness of the income tax and the different treatment of income stemming from work and from capital have been the main preoccupations of the authorities. Since a greater priority is given to domestic savings, dividends and returns on fixed-income securities were excluded from the clause on personal accumulation in the Income Laws of 1965 and 1966. He stated that the developing countries are subject to structural and institutional obstacles that curtail the effectiveness of fiscal policy, which has as its goals the creation and long-term expansion of the economic and social overhead capital by means of public investment, monetary and price stability in the short run, the stimulation and appropriate channeling of private investment, the protection and strengthening of the balance of payments, and the redistribution of the purchasing power of the population (1966, p. 9).

VII. *The External Sector*

Foreign Trade

Between 1945 and 1956, agricultural output displaced mining as the sector with the greatest value of exports. Since 1957, and especially in the present decade, the rate of growth of the agricultural sector and its exports has declined considerably. At the same time the current account deficit has steadily widened during the period of price stability.

The thesis about the external disequilibrium of Mexico, or the deterioration of the terms of trade, is frequently associated with the theory expounded by ECLA on the instability of external demand. The theory covers two aspects: the long-run trend of prices and the short-run fluctuations in the volume and value of exports. In general, the ECLA thesis on the deterioration of the terms of trade was accepted by the majority of Mexican economists, who recommended different solutions.

For Torres Gaitán, as well as many others, the decreases in the rate of growth of exports in the period 1950–1960 can be explained only by the weakening of external demand, rather than by internal factors, because of such unfavorable phenomena as the decline in prices which made the reduction of export taxes and the use of compensated trade operations ineffective (1961). In order to combat this instability, there must be international economic cooperation. World market conditions are very different today from what they were when Great Britain was the creditor center of world trade. Now that

this role has been assumed by the United States, it is being performed by a nation in which foreign trade represents a very small proportion of national income. Therefore, the United States has become a monopolist as a seller and a monopsonist as buyer (1954). Faced with this situation, Latin America turns out to be a marginal seller which does not affect the international position of that country. Torres Gaitán adds, however, that the trade policy of the United States makes the instability of demand more pronounced, for when the prices of raw materials decline, the United States argues that this is a market problem, but when there are price increases, the United States finds appropriate mechanisms to contain the rise of prices. Finally, international economic cooperation is in conflict with a vicious circle related to the fact that external financing is available if there is a development program. But such a program is not possible under conditions of an unstable foreign demand and the fluctuations of external prices (1954). In contrast García Reynoso (1963) maintains that there are no unsurmountable external obstacles to expansion arising from the different situations that occur in the markets for raw materials and those for manufactured goods, and that there is a tendency toward the expansion of trade in manufactures. Moreover, there are large markets that have not been exploited because of a lack of promotion and of sales technicians, especially with trading blocks such as the Latin American Free Trade Association and the European Common Market.

According to ECLA (United Nations, 1957) the solution of the problem depends on internal factors, such as the development of demand and substitution for imports. In other words, it implicitly accepts the conclusion that conditions in the international market cannot change favorably and that increasing use should be made of a protectionist policy.

Torres Gaitán was of the opinion that between 1950 and 1960 the rate of growth of imports was similar to that of national income, from which he inferred that the average propensity to import was stable. The growth of imports was parallel to that of national income, despite the increase in the price of imported goods, as well as higher import taxes and greater import restrictions. Moreover, imports expanded at a higher rate than exports to the United States. On the basis of these observations, Torres Gaitán concluded that the demand for imports is inelastic with respect to changes in external prices —an indication of the inflexibility of domestic supply—for a good part of total demand is oriented toward the foreign sector. He explains the deficit on current account in terms of the productive apparatus. According to him, there are a number or reasons for this. One of them is the higher rate of growth of income in Mexico than in the United States. Another is a constant propensity to import due to the inelasticity of domestic supply, a factor that acts together with the obstacles to exports. A third is the difference in the levels of productivity in the two countries which tends to make for a greater volume of Mexican purchases in the United States (1961).

Izquierdo provides other observations relative to the composition of imports. He regards the predominance of capital goods within imports as a vulnerable point in the economy of the country, because of the dependence of national industry on foreign sources of supply for equipment. He also points out that purchases by Mexicans in American border cities really constitute imports, even though they are counted as Mexican tourist expenditures. The large proportion of consumption goods included under this category somewhat changes the composition of imports (1964).

Some general conclusions can be de-

duced from the characteristics of foreign trade. Torres Gaitán, for example, points out that Mexico has had to finance its foreign trade with capital provided by the private sector and international organizations. He also infers that there is a hidden deficit in the balance of payments which is deferred by means of trade restrictions, although to date, no attempt has been made to correct it (1962). Mexico should increase its payments capacity in order to balance its financial relation at higher levels of output and trade, as required by development (1954). At this point, a change can be detected in Torres Gaitán's thought. In 1954, he asserted that there was no possibility of growth on the basis of an increase in exports greater than that of population, for the prospects of an increase in international demand were nil. Thus, he concluded that development must be based on industrialization, although, he added, it would be more practical to increase foreign exchange earnings, rather than to substitute for imports or receive foreign economic aid. A glance at his views on exports, written eight years later, reveals a different view with regard to the outlook for exports, and the same thing is true about his opinion of foreign aid. In 1954, he did not believe in it, but by 1962, he considered it decisive to finance development.

The proposed solution to the deficit on current account still falls within the realm of a protectionist policy. No economist seems to be in favor of a more liberal import policy to facilitate the exportation of manufactured goods.

Torres Gaitán describes the general objectives of Mexican trade policy. One of them has been to reach the maximum rate of development compatible with an equilibrium of the balance of payments. Another has been to increase the volume of exports with the highest degree of local manufacture. Others include limiting imports to the level of the capacity for foreign payments and improving the terms of trade through the diversification of markets and products (1962). Little has been written about the results attained by trade policy, although in general, the opinions expressed have been favorable. It has been pointed out that two tools have been used repeatedly—tariffs and import licences or quantitative controls. Of the former, it has been alleged that the average incidence of tariffs is very low—in fact it is even said to be one of the lowest in the world—although the arithmetic average used does not constitute an adequate criterion to evaluate effective protection. Another characteristic of Mexican trade policy which prevailed for several years was the decision to abstain from signing international trade agreements, such as the General Agreement on Tariffs and Trade, which might limit the country's freedom of action. However, this policy has changed to a certain extent since the Treaty of Montevideo was signed, creating the Latin America Free Trade Association.

A viewpoint held in common by such authors as Torres Gaitán (1954), Navarrete (1957), and Uriquidi (1962), is that they consider that the United States, which is a creditor in international trade, should abolish trade restrictions, while the less developed countries—and in Latin America, the countries affiliated with the Latin American Free Trade Association—should maintain their present restrictions against the industrialized countries. The justification for this position lies in the concept of trading power. In the creditor countries, foreign trade plays such a small role within the gross national product, that the elimination of trade barriers would not have a significant effect on their level of economic activity. On the other hand, the abolition of trade restrictions in the underdeveloped countries does not increase the purchasing power of the debtor countries, which actually depends on their exports. Far from producing a decline in

foreign trade, the requested concessions would have the opposite effect, for the development of the underdeveloped nations represents an increased income, and consequently, a greater volume of trade. The thesis is an attempt to explain the role of foreign trade within the economy of a developing country and to introduce an appraisal of protectionism. Nevertheless, it should be noted that according to its defenders, protectionism is valid only in the trade relations between a creditor nation and its debtor, less-developed trading partners, that is, it will not help to orient the trade relations among the latter.

Tourism

There is virtually no disagreement about the beneficial role of tourism within the Mexican economy. Two important functions have been attributed to it. The first is that it has been important in financing development, for it has been a source of foreign exchange that has paid for more than one-third of all imports. Secondly, it has been an offsetting factor in the downward trend of export prices, maintaining the capacity to import (E. Perez Lopez, 1961).

Nevertheless, Torres Gaitán (1954, 1961) remarks that the rate of growth of tourist expenditures in Mexico and that of tourist expenditures abroad constitutes a potential problem. Mexican tourism abroad is increasing at a faster rate than the inflow of foreign tourists, which might accentuate the growth of the foreign debt.

Latin American Economic Integration

The topic of Latin American economic integration was of great interest toward the end of the decade of the 1950s, at which time many works were written on the subject. The motives that led to the objective of integration, besides the example set by the European Common Market, can be classified as external and internal, although the former were perhaps more important. In general terms, the predominant external factor was the lack of confidence in exports, the fluctuations in the terms of trade and the smallness of the domestic market. The most frequently repeated argument to justify and promote the integration of Latin America has been the need to continue the process of industrialization, once substitution for imports has become more difficult.

The first topic to be discussed below will be the framework of international trade within which the country's development is taking place. Secondly, the process of integration will be considered, from the time Mexico joined the Latin American Free Trade Association (1962) through the mid-1960s. Finally, opinions on the prospects of integration and relevant recommendations will be presented.

It is a generally agreed fact that the share of Latin American exports within international trade has declined. On the other hand, as a result of industrialization, capital goods represent a higher proportion of imports, which increases the rigidity of imports and the dependence of the region on other countries in terms of production goods, which, as O. Campos Salas says, are more difficult to substitute for (1959). R. Vernon had pointed out earlier that the lag in the exports of Mexico and other countries was due to the decline in the prices of raw materials (1962, p. 528).

Urquidi upholds the point of view that protectionism is asymmetrical. Thus, the protection of agricultural products of the developed countries greatly affects the underdeveloped nations, while the protection granted by the latter to domestic production of the manufactures exported by the former does not affect the developed nations (1962, p. 422). The most self-sufficient industrialized countries

tend to bypass the underdeveloped countries. The underdeveloped nations, however, faced by an increasingly weak demand for their products, are unable to procure for themselves the capital goods they need. Because of this, Urquidi claims that world trade changes its structure and practices, and that the trend is toward greater protection, albeit among trading blocks. In effect, there has been a weakening of the General Agreement on Tariffs and Trade (GATT). To Urquidi, integration constitutes a means of attaining greater stability or more commercial security. On the other hand, it would also permit a better use of productive capacity, a problem that is more evident in the underdeveloped than in the developed countries, such as those of the European Common Market (1962).

According to Campos Salas (1959), almost all consumer goods are produced in Latin America, but in insufficient quantities. Hence, intraregional trade is an important tool for the creation of new market stimuli. In this way, integration would lead to an accelerated substitution for imports which would leave foreign exchange free for the acquisition of capital goods. The integration of Latin America, then is sought because of its effect in diverting trade.

Vernon (1962), who took part in a seminar on the subject, disagrees with the above proposition. He admits that world trade has changed and that the emerging nations need more stability, but the trend should be not to force import substitution but to reduce protection. He points out that GATT has made the reduction of tariffs possible, so that at present there are very few tariffs on manufactured goods, even though the opposite is true of agricultural commodities, for every country has wanted to raise the income of its rural sector. Even so, for the developing countries that have taken part in GATT, this has meant a means of obtaining new markets without reducing tariffs. But, he concludes, for Latin America and Mexico, the initial efforts at expanding exports have consisted in approaching partially protected markets, such as those of the Latin American Free Trade Association (LAFTA) (1962, pp. 528–529). LAFTA has brought about greater competition and a certain liberalization of Latin American trade among non–traditional industries. Wionczek (1966a) infers from this that there is a need to improve the present patterns of industrialization for Mexico and to think of them in terms of the entire region. While at the beginning businessmen were opposed to the liberalization of trade, today they are its staunchest supporters. Wionzcek wonders whether this is the result of the modernization of Latin America or of the need for new markets. He remarks, however, that every national authority acts unilaterally and generally outside the integration program, and that since the Treaty of Montevideo there have been no advances regarding the problems of international payments that result from trade.[8] Wionczek states that the slowness of the negotiations is not due exclusively to economic factors. The process of integration in Latin America is hampered by the disparity of the levels of development, because the industrialization programs entail an element of prestige—a motive that was overlooked in Europe. Furthermore, LAFTA does not have a mechanism for the industrialization of the less developed of the member countries. The repercussions of industrialization on the level of employment could be damaging, especially since the prevalent situation is one of underemployment; similarly, the State

[8] Wionczek is also of the opinion that the lack of progress in monetary cooperation stems from the many interpretations of the Treaty, for the monetary authorities are absorbed in the short–run problems, the plans are ambitious, there is a lack of experience, and many refuse to define their position.

as such, is fairly inefficient. The barriers to the expansion of growth are greater than in western Europe. And finally, tariffs, which were a protectionist measure at first, have become a source of revenue. Economic integration is also subject to the trade and aid policies of the developed countries. LAFTA has not yet reached the point of no return; political decisions are required to make it progress, but the larger Latin American countries have shown a limited willingness (1966a). Thus, there is also a political aspect which Wionczek regards as decisive for the establishment of concrete measures and that is the willingness and determination of the governments to participate actively in the process of integration, with a view to the long run. The coordination of the domestic policies of every country, Wionczek repeats insistently, must supplement that which is asked of the developed nations with regard to their attitudes toward LAFTA (1966b).

Other writers, more concerned with concrete changes, consider that the only solution lies in a process of gradual transformation—a period of transition—in order to reach a definite economic integration. According to Campos Salas, the period of gradual changes should take place under a Board of Directors and an Executive Committee of the Latin American Common Market, and he recommended a flexible schedule for 10 or 15 years. Among the most important items would be annual negotiations for tariff reductions and the unification of trade policy; equal treatment of investment coming from third countries, and a program to avoid double taxation (1959). A mechanism of this sort would be propitious for the adaptation of countries with different degrees of industralization to a system that would avoid the competition of non-homogeneous regions. As a general rule, the process would stress the principle of reciprocity. Countries would enter and remain in the Latin American Common Market only if they obtained an equitable part of the benefits (Campos Salas, 1959).

Generally speaking, the essential justification of Latin American economic integration is that regional industrialization would overcome the smallness of the market. In accordance with the present situation, the prevalent idea is that integration would be a most important stimulus for industrial expansion. This current of opinion is related to concepts mentioned above concerning the stagnation of exports of primary products and the difficulties involved in expanding those of manufactured goods.[9] On the other hand, there are the shared fears about the insufficiency of the domestic market. In sum, the basis of the idea of integration lies in the enlargement of domestic markets. Campos Salas comments on the phenomenon by saying that the new markets and free competition could lead to new poles of growth and the transmission of development to those places with appropriate conditions in terms of production factors, markets and transportation. Another advantage would be that regional output would absorb many raw materials that are exported today, with a subsequent positive effect on the world prices of such commodities—making them higher and more stable—and on the economic stability of the countries of the area (Campos Salas, 1959).

On the other hand, imports from third countries would not decline because of the increase in income that would result from industrialization, and a more rational use would be made of foreign exchange. Cam-

[9] In speaking of the outlook for Mexican foreign trade, Vernon summarizes the general consensus about exports. It is necessary to be aware that the expansion of exports of raw materials will not be fostered by favorable conditions in the near future, which means that all efforts should be oriented toward the exports of manufactured goods (1962).

pos Salas also foresees another important consequence of the rationalization of industrial production, and that is that plants would acquire optimum technological dimensions and economies of scale would bring a decrease in the costs of production. To this should be added a reduction in unutilized capacity, which is now sizable in certain consumer goods industries (1959). Wionczek introduces other considerations about the future performance of the process of integration when he calls for the improvement of the agreements on complementary industrialization and the execution of technical and financial assistence programs (1966b). A final point concerns the treatment of foreign investment: both feel that the establishment of basic guidelines is necessary. According to Campos Salas, this would prevent the advantages of integration from being absorbed by large international cartels. Thus, there would be a unification of policy to control dumping and subsidies to exports.

With regard to tariffs, the authors are in agreement when they call for the adoption of a common tariff and identical commercial and tax regulations (Campos Salas, 1959; Wionczek, 1966b).

The unification of monetary policy is another topic that has not been neglected. Uniformity of policies with respect to price stability, exchange rates, foreign trade credits, industrial development and wages should be promoted by the governments of member countries. From this point of view, Campos Salas is in favor of aid from the developed to the less developed countries by means of commercial credits and industrial promotion (1959).

In 1963, Wionczek called for giving priority to a regional monetary mechanism, because any rudimentary clearinghouse system would provide a saving of foreign exchange. In this way, it would be possible to achieve greater ease in payments, an improvement in the utilization of foreign exchange, and the creation of an institutional framework of permanent cooperation advancing toward a system for the settlement of multilateral payments (1966b).

VIII. *Economic Development*

While economic development has been discussed extensively, there are only a few fundamental ideas on the subject, accompanied by numerous digressions which take the form of small variations on central themes. On the other hand, it is necessary to recapitulate concepts which have already been mentioned, in order to make a more or less comprehensive presentation of the topic.

It can be said, in broad terms, that there are two predominant currents of thought about the process of development. There are those who insist on considering supply as the predominant factor for the whole economic system and others who consider the distribution and consumption factors as the motivating forces. Mrs. Navarrete, while supporting the former view, also takes the latter elements into account. Thus, she states that changes in the amounts and the combination of productive factors are the prime motivating force in the process. The changes which have direct effects are the changes in technology, the capital stock and the productivity of labor (1955, p. 231). Acknowledging that each developing country cannot be considered as an isolated system, Mrs. Navarrete claims that inward growth is subject to some limitations that arise from international relations, since the accumulation of capital is only possible with imported investment goods. Furthermore, the technology—which is also imported—is not always the most appropriate. Her recommendations about economic policy echo the ideas expressed previously; in order to sustain an endogenous growth, eco-

nomic policy should be oriented toward the application of technology to production, the increase of the stock of capital goods, and the increased productivity of labor (1955, p. 232).

The other school, which will only be mentioned in passing, is one already seen in the sections on domestic demand and the distribution of income. Its thesis could be summarized as that the impetus for development can arise only by means of the redistribution of income and the subsequent increase of the domestic market. The thesis is frequently laced with support for state intervention for the achievement of development. The viewpoints that are most frequently repeated are liberal, Marxist or of a third kind, which calls for a mixed economy. The latter is the predominant one among Mexican economists. In Mexico, state intervention is generally looked upon favorably. It is said that it is based on the principle of collective well-being. Emilio Mugica (1963) cites the Constitution of 1917 as the origin of the responsibility of the State to bring about structural changes, mainly through public investment. He believes that State intervention has been compatible with private enterprise and that there is no single, rigid criterion for the determination of the functions or size of public institutions. This position supports the pragmatic and flexible approach that has been characteristic of Mexican economic policy. More specifically, Mrs. Navarrete recommends the intervention of the State in key sectors which are complementary to each other and where intervention would permit a rapid increase of income, a principle which is akin to the theories of Rosenstein–Rodan and Hirschman on unbalanced development in strategic sectors. In the case of Mexico, she notes, this is true in certain export crops and mass consumption industries, such as textiles and foodstuffs. Similarly, she is of the opinion that the State must intervene as an entrepreneur–innovator and as a banker. Nevertheless, as a capital market is formed, public financing will be gradually replaced by private financing, even though there are still fields that have to depend on government financing (1955, p. 233).

Besides adopting a similar position on State intervention within the process of development, leftist writers adopt criteria of an international political nature. Carmona (1963) claims for example, that underdevelopment is not the result of a scarcity of capital, but of the political and economic subordination imposed by colonialism.

The financing of development is a subject that has concerned many authors. Among them, Sáenz is a clear exponent of the traditional current. According to him, the main thing is to decide on the amount of resources that will be earmarked for capital formation and subsequently how its cost will be distributed. The income of the country goes into consumption or investment; for this reason, consumption can be increased only to the extent that there are reductions in investment in development programs or that use is made of the monetary reserves and the holdings of foreign exchange (1958, p. 624). Moreover, in the short run, it is necessary to take into account the inelasticity in the supply of many production goods, and that wage increases without a concurrent and proportional increase in productivity are far from being a permanent solution to the social problem of the rise in prices. They would only amount to a temporary redistribution of income, in favor of some groups and to the detriment of others whose money incomes do not rise in the same proportion.[10] Economic policy as

[10] When there are price increases, the solution is to reduce the volume of the money stock to a level that is congruent with the amounts of goods available for consumption. If there are problems because of inflation the solution can take one of three forms: an increase in voluntary saving, forced saving because of price rises, or increases in overall taxation (Sáenz,

Sáenz points out, should seek the allocation of resources, distributing the cost equitably within the society, without affecting exchange stability. The forms that these measures can take are well known, and the most important among them occur in the tax system, price changes, stimuli for saving, credit and finance facilities for selected purposes, and changes in the creation of money by the government (1958). On the other hand, Sáenz maintains that the possibility of capital formation by means of a monetary expansion and stimulating saving is limited by the need to keep the balance of payments in equilibrium and because of the inherent contradictions between monetary expansion and the growth of voluntary saving. In fact, he holds the view that stability within development depends basically on changes in the tax system and on maintaining short–run equilibrium. He also claims that the financial and fiscal systems are inadequate for the control of price increases. He feels that there is a generally accepted assumption that the country can invest and reinvest an increasing proportion of its resources without reducing total consumption. In reality, however, price rises and devaluations have been a direct means of holding down total consumption to a level compatible with the level of actual public investment (1958, p. 630).

Navarrete (1956) considered that economic policy had been based on credit and trade policies, as well as public spending policy. But tax policy and financial policy (in the securities market) have yet to be developed. He attributes the price increases of the two decades prior to 1957 to the excess of investment in relation to saving. Excessive investment is not harmful in itself, for it is essential for developing economies. The harm usually stems from the means used to obtain it. At the time, Navarrete already felt that there had been a reduction of the excess of investment over saving in the public sector. Similarly, he favored the tax system as the source of financing for public spending.

The Process of Development in Mexico

There are divergent opinions on the form taken by the development process of Mexico. In order to organize them, a summary will be presented first of those viewpoints based on the concept of balanced growth, and in the following section an examination will be made of those that consider economic policy from the standpoint of the measures taken to implement growth.

With regard to output, the opinion held by Urquidi (1961) is of special interest, for he considers that the growth of the productive apparatus has been balanced in the past few years. The increase of output has been abetted to an important extent by technological innovation as well as social overhead capital projects executed by the public sector. Nevertheless, there have been some obstacles, such as the coefficient of capitalization, which has fluctuated between 10 and 15 percent, a rate considered by Urquidi to be insufficient for the requirements of the country. (In 1970, it was over 20 percent.) Concerning industrial growth in particular, he maintains that this phenomenon has permitted the absorption of excess rural population and an increase in the income of the urban sector. A similar opinion has been expressed by Fernandez Hurtado (1960), who also believes that the increasing output of durable goods and capital equipment is an indication of a new stage in the process of industrialization in Mexico. Urquidi still has some reservations about it. While it is true that industrialization provides a stimulus for commercial agriculture and helps to maintain the balance of payments in equilibrium through the substitution for im-

1958, p. 626). The second alternate would be the one to follow if the other two are not operative.

ports, the rate of industrial growth has been low. His statement is based on criteria such as external equilibrium and the ability to solve the problem of absorbing surplus labor, neither of which has been attained by development. The roots of the slowness of growth lie in the smallness of the market—a statement of dubious validity repeated continually—as well as in the lack of integration in the process of development and in the demand for imported durable consumer goods. He also considers the subsector of intermediate goods as the most vulnerable point of the industrial structure. Regarding the agricultural sector, Urquidi subscribes to structuralist concepts about the rigidity of agricultural supply, but adds that the most dynamic segments of agriculture are those which are tied to foreign markets (1961).

What solutions have been proposed? Those advanced by Urquidi are important because they encompass the opinions of other authors as well. With respect to industrialization, he recommends that decision-making be based on two criteria—external equilibrium and the prevention of unemployment. (To these, García Reynoso adds considerations of costs and location). Expansion should be vertical as well as horizontal. Industrial integration, an important objective in recent years, makes planning and its enforcement necessary in all sectors. The acceleration of the production of intermediate goods will be to the advantage of implementation of the two previous recommendations, because of their large share within exports. Moreover, the existence of a steel industry would permit an increased output of capital goods.

IX. *Attempts at Economic Planning*

In Mexico, economic planning, in a technical sense, has never gone beyond the level of well-intentioned attempts. The first attempts at planning date from 1953, when the Investment Committee, an office which was directly affiliated to the Presidency of the Republic, was created to determine the priority of different public investment projects. In 1959, the Committee became a ministerial office. After a decline in business activity in 1962, an Intersecretarial Planning Commission was created, which put into effect a Plan for Immediate Action, which defined levels of public investment and related projects. In 1965, with the evaluation of the results of this plan, another was prepared for the period 1967–1970. By 1963, a number of works had appeared which commented on the characteristics of the Planning commission, its results, its future potential, and its role in the development of the country. At the time, many opinions were expounded, which will be divided into two groups and presented below. The first group appraises the experiences of the country in terms of planning. The second examines the prospects of planning and how it could be improved.

In the first place, it is important to establish just what kind of planning has been done in Mexico. Writers on the subject agree that it has been of an indicative nature, that is, free of dictates even within the public sector. Such is the standpoint of G. Esteva (1963), and J. Zamora Batiz (1967). De la Peña, on the other hand, considers that it has been a mixture of a centralized type of planning with pluralistic characteristics, although it depends to a great extent on the ups and downs of politics. The predominant criterion has been to associate investment and development, at least, implicitly. But different approaches, such as relating productivity and development or employment and development, would also be valid (1963). This situation is not unusual, considering the prevalent current of thought among the experts on the subject, who feel that development is directly related to investment.

Some critics consider that the results

have not been altogether favorable. Using a strict definition of planning, Esteva states that it does not exist yet in Mexico. But de la Peña says only that it has not been sufficient, for long-term goals there have not been established and frequently the means are confused with the ends. According to Esteva, the well-known Plan for Immediate Action was insufficient as an economic development plan because it was formulated before a long-term plan had been completed, because it was essentially a conglomeration of individual investment projects which did not adhere to aggregate requirements, and because, despite its indicative nature, the private sector did not take part in its formulation. He believes, however, that it has been useful for future programming of public spending and in fostering an awareness of planning (1963). The two latter ideas are shared by De la Peña and Zamora Batiz. The latter mentions other results of the first steps taken in planning, such as the creation of an awareness of problems like external financing and investment in self-financing projects, which have awakened a positive interest within the private sector (1967). But these writers also point out drawbacks, some of which are linked to institutional factors. A far-reaching problem with regard to planning projects has been brought out and criticized by De la Peña (1963), who asserts that the planning project has not been realistic enough, because it does not establish structural changes in the society nor in public administration. There is ignorance with respect to the operation of the public sector. Another problem that cannot be justified, which he points out, is that the Planning Commission does not have executive functions and hence, because of a lack of authority, it becomes a mere advisory body, which detracts from the willingness to plan. The same author notes that there are errors due to false extrapolations that result from the fact that the plan does not conform to public administration conditions nor to the social structure of the country (1963).

There are some explanations of the shortcomings of or the obstacles to attempts at planning. Sharing the ideas of I. Pichardo Pagasa (1964), and J. Zamora (1967), S. De la Peña makes note of a series of factors that limit the possibilities of planning, for example, the lack of statistics, insufficient research, the scarcity of technicians, and of a lack of technical criteria for sectorial programming. He insists that frequently the means are confused with the ends, so that state intervention depends on the individuals in power. Authorities at a lower level do not take up a position that would be supplementary to the plan. And last but not least, there is no appropriate short-run planning mechanism.

The experiences under discussion have led the same authors to make recommendations for a more effective system of planning. There is agreement that the process of the formulation of the plan should be more rational and that it should not lose its indicative nature. S. De la Peña maintains that a plan should consider two distinct aspects—the "planning of planning" and the structure of the methods to be applied (1960).

X. *Economic Policy, Economists and Mexican Nationalism*

A review of the literature published on the economic policy of Mexico reveals surprising facts and many inconsistencies. (In this section, I will refer to the common denominator of published works, not to those of outstanding, well-prepared economists whose ability is unusual and who undoubtedly have been quoted in the course of this work.) Mexican economic policy has not been examined systematically. At best, it has been examined only partially, for the formulation of economic policy in Mexico is so closely tied

to the goals of the Revolution and to nationalism that it is not easy to separate them. Yet, by going to the bottom of it, one can see the various interests at play, the ideological positions and the price that the community must pay for them. In this final section, general comments will be made on the analysis of economic policy reviewed in this work, some points will be touched upon that are not discussed or analyzed by Mexican economists and an attempt will be made to examine and explain the actual practice of economics in Mexico.

Three main points stand out when one reviews the literature by Mexican economists on economic policy. The first is the nearly total absence, until recently, of the analytical tools of modern mathematical economics and general economic analysis, including the field of economic policy. At times this leads to misunderstandings and confusion on controversial matters. Sometimes a concept which expresses a mathematical relationship is given a different meaning by different people due to the lack of a precise definition, or on the contrary, different terms are used for the same concept. Frequently, descriptive charts and graphs are used to illustrate hypotheses which are never proved with statistics. In fact, there is a great deal of mistrust and suspicion connected with the use of quantitative methods.

The second point that should be emphasized is the lack of interest in abstract thought. At times the object of analysis is too general and comprehensive, and that which is gained in incidental and general terms is lost in analytical shallowness of treatment of the main theme. Of course, the more variables are included in the analysis, the clearer the need for using mathematical methods to be able to handle the multiple interrelations. All this leads to a lack of precision with respect to the assumptions of the analysis, which more often than not are not made explicit. This becomes more evident when various fields are examined by an author and his assumptions inadvertently seem to change, so as to keep coherence with his stated conclusions. This obviously makes it difficult to follow the logic of the analysis and, moreover, lends itself to never-ending discussions since, as mentioned above, the merits of different works of analysis are based upon conclusions that have been derived from different assumptions, without keeping in mind the diversity of the latter.

The third point that should be brought out is that some areas have not been analyzed and yet are of substantial interest to economists in other countries. When dealt with, these points are described in a routine manner with no critical analysis. Both aspects are combined in the position of the economist in Mexico, his employment almost exclusively in the public sector, his tendency to become involved in politics, and above all his being under the influence of the nationalism which permeates Mexico. The final comments will be devoted to an explanation of the situation.

Harry G. Johnson (1967) has analyzed nationalism in recently-formed states. It is interesting to compare his observations with Mexican nationalism and with the recommendations of the economists dealt with in this paper. Johnson maintains that nationalism implies an ideological preference in economic policy for a number of goals. One of these goals is self-sufficiency; another is the entry of public property into strategic sectors of the economy, or where this would be impractical, to replace it with a broad regulation and control of private firms.

In short, nationalism contributes certain specific traits to economic policy, the most important of which is the insistence on industrialization that sometimes results

in the relative neglect of agriculture and often in the deliberate exploitation of the agricultural sector in order to finance industrialization. A second characteristic that can be attributed to nationalism is the preference for economic planning. A third trait is the indiscriminate hostility towards the big international corporations. A look at Mexican nationalism is now in order.

Actually, there are two distinct phases in the economic nationalism produced by the Revolution of 1910. The first of these covers the period until 1940; the second has continued to the present day. The latter part of the first phase coincides with its most constructive period—fiscal policy was introduced as a tool to foster economic development. Defense expenditures became less important, while education acquired great importance, as did communications and irrigation projects. With the oil expropriation, foreign firms were unquestionably subjected to Mexican laws. The distribution of lands gathered force and the political power of the large landowners (*hacendados*) was abolished. However since World War II, the Government has adopted an ideology that favors economic growth, albeit in a different way. This ideology comprises all social groups and acquires a definite form—national interest, national unity, general submission to the State, xenophobia, and the casting into oblivion of the class struggle. Thus, the nationalist ideology which embraces economic development as the goal of the Revolution was established and there an attempt was made by the Party to unite all of those taking part in the economic process. For their part, the middle and upper classes ceased to be hostile to the government and began to participate in the political life of the country. At the same time, as industrialization became the main object of economic policy, gradually the government listened more and more attentively to the entrepreneurs, and public officials wielded their power by passing judgment on each individual case without applying any general pattern of criteria. Although this system undoubtedly created uncertainty and constituted a risk, the decisions were usually favorable to production, investment and profits, and it established a balance, as well as a common understanding between government officials and entrepreneurs. Thus, the dynamics of growth, imparted by the private sector, were facilitated by avoiding the stumbling blocks inherent in labor relations problems.

As a result of the industrialization policy, there was also a bias in the distribution of the national income. Profits increased by means of higher prices for manufactured goods, while the price of agricultural inputs, those supplied by the public sector and the wage levels remained stable. In effect, the policy of agricultural and industrial development led to a redistribution of income, in terms of payments to the factors of production, which favored the middle class, especially the cultured middle class and the upper class which owned the means of production, and hence represented the means for the formation of capital.

From a social point of view, the nationalism that arose from the Revolution had two extremely positive effects. The first of these was the elimination of the previous social stratification. The second was the formation of a new social and economic structure that was modern and capable of adjusting to economic development, as evidenced by greater factor mobility and by the fact that the social system can assimilate changes without destroying the mechanism of collective harmony (M. Nash, 1967). Both aspects were propitious for the continuous growth of the national product.

The nationalist policy gave employ-

ment, energy, and a purpose to a great number of dispossessed Mexicans. At the same time, it made it possible to obtain immediate returns by employing the existing factors of production more intensively, especially, because of the agrarian reform, land and labor. Similarly, demographic mobility and the distribution of land in the form of *ejidos* kept farmers from restless ferment. The increasing levels of urban employment, the paternalism of the State, and the increase in the supply of labor made the labor unions more ductile. Meanwhile, the process of economic development transferred a larger share of the increases in income toward the middle and upper classes in the form of returns to education, oligopolic profits and other payments to property.

Nationalism created a consensus of opinion in favor of the Mexican identity and simultaneously transformed the social structure and the set of values. (In the Mexican case there is also another type of national identification, namely the *indian* and *mestizo* cultural heritage.) It also played a role in economic decision-making and helped to keep the political process outside and prevent it from disturbing the economic process, with the qualification that once the ends were accepted, the means became untouchable and unquestionable. Nationalism, in the wake of patriotism and the Mexican identity, made it more and more difficult to pass judgment on political centralism, paternalism and economic policy; in effect, the public sector is immune from independent criticism. The same is true of the distortions created by economic policy. To some extent, to exalt nationalistic values is a way of preventing criticism, maintaining privileges and defending vested interests.

What has been the role of the economists in the formulation and implementation of economic policy? The formal study of economics in Mexico began in the 30's within the Law School of the National University. It was the outgrowth of the need to understand and guide the increasingly complex economic system produced by the Revolution. From the Law School it inherited the humanistic tradition which became linked to a Marxist ideology and the rejection of mathematics and quantitative methods, that it has retained until now. The other schools of economics, founded later on, have imitated the curricula of the former and are still rather small or too recent to make their presence felt yet.

The economists, like the rest of the intellectuals, are identified with the Government and are widely employed within it. At the same time, like all who take part in politics, they subscribe to all of the dicta of the Revolution. However, the very fact that they work for the Government, together with the evidence of economic development, has limited the critical approach of Mexican economists. They are frequently reduced to the justification and rationalization of economic measures that government officials and politicians have taken without consulting them. For the most part, economists, like the rest of the intellectuals, have lent their support to the economic policy that has been followed, despite the fact that it provides measures—especially with regard to imports and investment incentives—which are partial to the already considerable privileges of the industrialists. Similarly to most of the left, they share these tenets with the so-called national bourgeoisie. For example, together with the members of the National Chamber of Manufacturing Industry (*Cámara Nacional de la Industria de Transformación*), they agree with industrialization and its promotion by means of import controls and tariffs; they also agree on the position taken on limiting foreign investment. These econo-

mists also maintain some objectives in common with the Confederation of Industrial Chambers (*Confederación de Cámaras Industriales*), although not all, for they differ on foreign investment, which the Confederation considers beneficial. Such a position is logical from the standpoint of the industrialists, but not from that of the nationalistic left, which is upheld by most economists. It is well known that these economists extol a better distribution of income, which is incompatible with and contradictory to the type of inwardly-oriented industrialization, based on import substitution, import controls and oligopolistic profits that obviously distort the distribution of income. Actually, in this respect they play up to the bourgeoisie. This incongruency can be explained as a reflection of the absolute distrust of these economists for market forces and their support of economic controls. There is also a tendency for economists to support State intervention in any form, and a partiality to production. However, in the last case, the transfer of income that results from an increase in the price of manufactured goods implies a bias favorable to industrial profits.[11]

The economists have adopted these positions as part of an ideological attitude and because the leftist position is extremely attractive, in that it gives those who uphold it a sense of being endowed with a social ethic. It should be pointed out that when this position has been combined with analytical objectivity and technical competence, it has produced the best Mexican economists. But in many cases, the reason for supporting the left is the fear of the well-known epithets applied to certain attitudes. Nobody wants to be classified as a non-nationalist when that implies being a reactionary. It is much more attractive to side with the underdog and be against imperialism. It is also more convenient, in order to avoid social criticism, to adopt a nationalistic position without an objective evaluation of the means proposed for the attainment of the goals.

The economists of the extreme left, who are generally sheltered within academic work, have maintained a static position that distorts reality. It is less and less appropriate for the dynamic conditions of the country, because that outlook has not been modified in the face of the changes in the prevailing reality. A Latin American economist wrote, ". . . most of the practicing Marxists have done little more than to continue to repeat and uphold the work of the master, without trying to use his methodology to prove those areas which remained outside of his scope or to enrich the original model with the new ones introduced by historic evolution. . . . If this is a valid assertion . . . it is even more so with regard to Latin America" (Espartaco, 1964, p. 68). The Marxist influence in the formation of Mexican economists has contributed to limiting their criticism of economic policy, for the branch of economics which lends itself best for this analysis, welfare economics, applies the theory of utility—which is repudiated by most Marxists. Moreover, welfare economics has a mathematical base that is not easily understood by the ordinary, non-mathematical economist—as are most of those in Mexico—who finds it hard to grasp the collateral effects of the proposed measures. It is precisely the coherence that welfare economics gives to economic policy through the analysis of

[11] Undoubtedly, this was necessary in order to launch industrial development. Nevertheless, it has been shown that at present the industrial development of Mexico is not greater than the international norm, i.e., it has not advanced more than could be expected from a country of its size and output per capita. Moreover, between 1950 and 1960—when there was a decline in industrial development by international standards. In fact, comparatively speaking, the relative share of industry within the gross national product is smaller today than it was in 1950.

general equilibrium which permits a better understanding of the side effects, which cannot be observed with the casuistic analysis of partial equilibrium based on formal logic, practiced by these economists. Thus, it is understandable, for example, that economists support import controls and substitution for imports through domestic production in order to increase the domestic product and employment. In general, however, they do not discuss additional effects, such as the creation of oligopolistic profits—of which foreign investors come to take advantage—that distort the distribution of income in favor of the proprietor class. Moreover, they do not take into consideration the systematic increase in the price of inputs, which acts to the detriment of exports and makes it necessary to incur a foreign debt in order to maintain the rate of development. These results are very different from the nationalistic goals that inspired them.

In this respect, some examples might illustrate the usefulness of a revaluation of the future economic policy of the country, which, needless to say, could also be useful from a political standpoint. This implies passing judgment upon some of the dearest symbols of Mexican nationalism, analyzing their effects and going beyond paying them lip service—generally done to avoid political suicide. It also means that economic information should be disclosed to allow informed criticism, and ceasing to keep much information confidential, which practice in the final analysis acts in favor of personal interests.

Once on the road toward a market economy, in accordance with the Constitution, that is viable and dynamic, it would be self-defeating to hamper it with "revolutionary" or nationalistic attitudes which hinder factor mobility, monopolize goods and impede the transmission of stimuli arising from market incentives and economic forces. In the past, a greater factor mobility and a better utilization of resources led to higher rates of development. But even today, mobility is far from adequate for the attainment of greater efficiency and, in certain cases, the economy tends to lose flexibility rather than gain it. In agriculture, for example, the collective *ejido* proved to be more productive at first than the individual *ejido* or privately-owned farms. However, the collective form of organization threatened to create a political unit that would be difficult to control, and in order to avoid this, preference was given to the individual *ejido* (S. Eckstein, 1966). The latter form of production gave rise to a market composed of small farmers, imperfectly competitive, which divested the farmer of bargaining power vis-a-vis the jobbers and other middlemen with no limitations of property and transfer—as the *ejidatario* has—and who, free of all bonds, have evolved and become strong economically, creating a bourgeoisie in medium-sized and small cities that profits at the expense of the small farmer. Although in general the collective *ejido* seems to be preferable and it would be advantageous to expand it as much as possible and enable it to absorb new technology, the scant education and civic experience of the *ejidatarios* limit its performance. By the same token, there is a scarcity of technicians who could help them achieve an adequate organization and create mechanisms for the formation of capital. Efficiency and mobility could be supplemented more effectively than at present, if, in addition, a market for the lease of lands were permitted to develop. In an autonomous way, this market would amass lands into more efficient units which would increase output—a result that is obtained directly by means of the collective *ejido*. The producers would also have greater bargaining power in dealing with the agricultural

middleman. Yet the leasing of *ejido* lands is violently attacked because it would give rise to a new form of latifundia,[12] and because it is not "revolutionary," although it would seem that what really was not revolutionary was to support the individual *ejido*. The end result has been a clandestine market for the lease of *ejido* lands which does not facilitate sufficiently the creation of more productive units or the functioning of an agricultural sector oriented to the market and an adequate allocation of productive factors.

The same thing is true when import controls, in the form of protection given to domestic producers of the same goods, restrict the competitive conditions in industry and foster excessive oligopolistic profits. A gradual, greater liberalism in trade policy would force entrepreneurs to become more efficient. It would help the more competent among them, through lower input costs and a greater demand for goods, by facilitating the exportation of goods. It would accelerate technological change. It would make the processes more labor–intensive and less capital–intensive. It would lower consumer prices and allow the consumer to share the increases in productivity. It would also raise the real income of wage–earners, farmers and the middle class. However, in this case, the desire to have an industrial base as large as possible has led government officials, in some instances, to associate the people's welfare with the number of import commodities that require an import license.

In the case of natural resources, the traditional standpoint views the sale of the products of the subsoil almost in the same light as the sale of national territory. This prejudice helps to explain the decline of mineral exports in the recent period of exchange stability.

The desire to restrict commercial relations with the rest of the world to limited exports while proceeding with import substitution has led to the establishment of plants that are below optimal size, and that have operating costs that are far above the prices on the world market. Despite the large size and comprehensiveness of the public sector, economists have not written about its magnitude, operations or efficiency except to express conventional compliments. One would think that the use of natural resources and the possibility of substitution for them is conditioned by the available technology—scientific progress and its adaptation to local conditions makes for a lesser dependence on natural resources, or makes them less essential. For this reason what a national or state–owned firm does with the physical resources is not so important as what it does with human beings. In other words, the State should ensure that people are allowed to get an education, to be free and capable men, to use the available technology, to master their environment or overcome the scarcity of some resources with the abundance of others. In this respect, it is clear that the Revolution has been successful.

In order to gain a better knowledge of economic policy, it is necessary to practice a constant and sincere self-criticism. A substantial and growing number of Mexican economists are returning to the country after doing graduate work abroad. New schools of economics founded on a modern basis are now graduating competent economists. Both types are well trained in modern techniques of economic analysis and are beginning to destroy old myths with an iconoclastic attitude. Moreover in the school of economics of the National University, former students of graduate schools of foreign universities

[12] Although it is evident that a contemporary commercial agricultural unit that is large, productive and efficient is very different from the old Porfirian latifundium, they are often taken to be synonymous. But this tends to happen when a static view is taken of economic phenomena.

are influencing students and reforming programs, exerting pressures so that the teaching of economics will be brought up to date and, generally, questioning the old and rigid educational systems that have constrained greater improvement within the profession.

References

A. Bueno Y. Urquidi, "La Iniciativa Privada y el Estado en la Promoción del Desarrollo Económico," *Actividad Económica en Latinoamerica*, June 15, 1962, 23, 5–8.

J. Campillo Sainz, "Tesis de la CONCAMIN sobre Inversiones Extranjeras," *Actividad Económica en Latinoamerica*, July, 1966, 72, 13–15.

O. Campos Salas, "Comercio Interlatinoamericano: Integración Regional," *Investigación Económica, ENE*, 1959, *XIX*, 76, 589–606.

F. Carmona, "Dependencia y Subdesarrollo Económico," *Investigación Económica*, ENE, 1963, *XXIII*, 90, 385–4230.

J. L. Ceceña, "Las Inversiones Extranjeras en México, "*Investigación Económica*, ENE April-June, 1965, *1*, 98, 271–99.

A. Cervantes Delgado, "La Política Fiscal y las Reformas Impositivas de 1962, "*El Trimestre Económico*, FCE, July-Sept., 1962, *XXIX*, 115, 391–409.

J. S. De Beers, "El Peso Mexicano 1941–1949," *Problems Agricolas e Industriales de México*, 1953, *5*, 1, 7–134.

S. De La Peña, "Hacia la Planeación del Desarrollo en México," *Comercio Exterior*, 1960, 610–13.

———, "Sobre el Presunto Proyecto de Ley," *Comercio Exterior*, Nov. 1963, 819–820.

M. A. Duran, "Condiciones y Perspectivas de la Agriculturea Mexicana," *El Trimestre Económico*, FCE, Jan.-March, 1961, *XXVIII*, 109, 52–91.

———, "La Reforma Agraria Mexicana," *Comercio Exterior*, Jun. 1968, 493–98.

S. Eckstein, *El Ejido Colectivo en México*, FCE, México 1966.

———, *El Marco Macroeconómico del Problema Agrario Mexicano*, Centro de Investigaciones Agrarias, Comité Interamericano de Desarrollo Agrícola, Documento Preliminar, México, D. F. 1968.

Espartaco, "Crítica del Modela Político Económico de la Izquierda Oficial," *El Trimestre Económico*, FCE, Jan.-March, *XXI*, 121, 67–92.

G. Esteva, "El Mito de la Planeación Económica Mexicana," *Comercio Exterior*, Nov. 1963, 822–26.

E. Fernandez Hurtado, "La Iniciativa Privada y el Estado como Promotores del Desarrollo," in José A. Iturriaga, ed., *México, 50 Años de Revolución*, FCE, Tomo I, Mexico, 1960, pp. 595–619.

R. Fernandez y Fernandez, "Una Nueva Política Agraria," *Revista de Economía*, May 15, 1954, *XVII*, 145–48.

———, "La Reforma Agraria Mexicana: Logros y Problemas Derivados," *El Trimestre Económico*, FCE, April-June 1957, *XXIV*, 2, 143–159.

———, "Notas sobre el Problema Agrario Mexicano Actual," *El Trimestre Económico*, April-June 1960, *XXVII*, 106, 203–8.

H. Flores de la Peña, "La Mecanica de la Inflación," *Investigación Económica* ENE, 1953, *XIII*, 4, 461–481.

———, "Crecimiento Demográfico, Desarrollo Agrícola y Desarrollo Económico," *Investigación Económica*, ENE, 1954, *XIV*, 9, 519–35.

———, "Agricultura Mexicana," *Comercio Exterior*, July, 1958, 376–79.

H. Flores de la Peña and A. Ferrer, "Salarios Reales y Desarrollo Económico," *El Trimestre Económico*, FCE. Oct.-Dec. 1951, *XVIII*, 72, 617–28.

E. Flores, "La Localización de la Agricultura y los Cambios del Uso de la Tierra en México," *Boletín de Estudios Especiales*, Feb. 13, 1959, *XIII*, 151, 161–74.

———, *Tratado de Economia Agricola*, FCE, 3rd ed., Mexico. 1964.

———, "La Teoría Económica y la Tipología de la Reforma Agraria," *Comercio Exterior*, May 1967, 379–85.

P. Garcia Reynoso, "Veinticino Años de Política Mexicana de Comercio Exterior y sus Resultados," *Comercio Exterior*, July 1962, 406–09.

———, "Evolución de la Estructura de las Exportacion, Mexicanas," *Comercio Exterior,* Aug. 1963, 569–71.

———, "La Política Mexicana de Fomento Industrial," *Comercio Exterior,* Nov. 1968, 959–64.

R. Gomez, "Estabilidad y Desarrollo-El Caso de México," *Comercio Exterior,* Nov. 1964, 778–82.

———, "Evolución del Aparto Financiero Mexicano," *El Mercado de Valores,* March, 6, 1967. 10, 199–202.

H. G. Johnson, "The Ideology of Economic Policy in the New States," in *Economic Nationalism in old and New States,* Chicago, 1967, 124–42.

M. Hinojosa Ortiz, "Reflexiones sobre una Política Agraria," *Investigación Económica,* ENE, 1961, *XXI, 82, 211–234.*

———, "Posibilidades de una Reforma Fiscal en Materia de Impuestos sobre la Renta," *Revista Fiscal y Financiera,* April, 30, 1964, *XXIV,* 202, 35–50.

R. Hoyo D'Adonna, "El Sistema Fiscal en el Financiamiento Directo del Desarrollo," *Actividad Económica en Latinoamerica,* April 1968, 93, 5–6.

R. Izquiredo, "Protectionism en México," in R. Vernon ed., *Public Policy and Private Enterprise in Mexico,* Cambridge, Mass., 1964, pp. 241–89.

N. Kaldor, "¿Aprenderán a Gravar los Países Subdesarrollados?," *Comercio Exterior,* Jan. 1963, 46–8.

———, "Las Reformas al Sistema Fiscal en México," *Comparcio Exterior,* April 1964, 265–67.

G. Lira Porragas, "Desarrollo Agrícola vs. Desarrollo Industrial," *Revista de Economía,* 1964, *XXVII,* 5, 129–40.

E. Lobato Lopez, "La Políticia Monetaria Mexicana," *Investigación Económica,* ENE, 1968, *XXIII,* 72, 557–81.

F. Lopez Rosado, "La Situación Monetaria de México," *Investigación Económica,* ENE, 1953, *XIII,* 4, 441–54.

E. Mugica, "Participación del Sector Público en el Dosarrollo Económico de México," *Actividad Económica en Latinoamerica,* Oct. 1963, 39, 49–51.

M. Nash, "Economic Nationalism in Mexico," in Harry G. Johnson, ed., *Economic Nationalism in Old and New States,* Chicago 1967, pp. 71–84.

A. Navarrete, "Las Relaciones Financieras Internacionales de México," *Investigación Económica,* ENE, 1955, *XV,* 2, 179–89.

———, "Productividad, Ocupación y Desocupación en México: 1940–1965," *El Trimestre Económico,* FCE, Oct.-Dec. 1956, *XXIII,* 4, 415–33.

———, "El Sector Público en el Desarrollo Económico," *Investigación Económica,* ENE, 1957, *XVII,* 1, 43–61.

———, "El Crecimiento Económico de México y las Inversiones Extranjeras," *El Trimestre Económico,* Oct.-Dec. 1958, *XXV,* 100, 556–69.

———, "La Inversión Extranjera Directa en México," *El Mercado de Valores,* Oct. 31, 1966, 44, 1079–1183, 1104.

I. M. de Navarrete, "El Proceso de Desarrollo Económico y la Política Fiscal," *Investigación Económica,* ENE, 1955, *XV,* 2, 229–47.

———, "Política Fiscal y Desarrollo Industrial," *Comercio Exterior,* Jan. 1956, 26–8.

———, "La Política Fiscal y la Distribución del Ingreso," *Investigación Económica,* ENE, 1957, *XVII,* 1, 83–42.

———, "Notas sobre la Distribución del Ingreso Nacional en México," *Investigación Económica,* ENE, 1959, *XIX,* 73, 31–39.

———, "La Naturaleza de la Reforma Fiscal," *Comercio Exterior,* March 1962, 138–142.

———, "The Tax Structure and the Economic Development of Mexico," *Public Finance,* 1964, *XIX,* 2, 158–99.

A. Noriega Herrera, "Las Devaluaciones Monetarias de México, 1938–1954," *Investigación Económica,* ENE, 1955, *XV,* 1, 149–77.

J. Noyola, "¿Existe una Política de Precios?" *Revista de Economia,* Aug. 15, 1948, *XI,* 8, 9–10.

———, "El Desarrollo Económico y la Inflación en México y otros Países Latinoamericanos," *Investigación Económica,* ENE, 1956, *XVI,* 4, 603–6.

——— y D. G. Lopez Rosado, "Los Salarios Reales en México 1939–1950," *El Trimestre Económico*, FCE, April-June 1951, *XVIII*, 70, 201–09.

C. Ocaña Y J. Alejo, "I. La Carga Fiscal en México, su Evolución y su Importancia," *Comercio Exterior*, Feb. 1963a, 80–82.

——— y ———, "II. La Carga Fiscal en México: El Impuesto sobr la Renta," *Comercio Exterior*, Feb. 1963b, 169–73.

R. A. Ollervides, "La Nacionalización y la Inversión de Captales Extranjeros en la Industria," *Comercio Exterior*, July 1966, 486–91.

A. Ortiz Mena, "La Política Financiera en los últimos cinco años," *El Mercado de Valores*, Nov. 18, 1963, 46, 597–600, 605.

———, "Contenido y Avances de la Política Fiscal," *Actividad Económica en Latinoamerica*, Oct. 1966, 75, 5–18.

E. Padilla, "La Devaluatión del Peso Mexicano: Cuatro Conferrencias," *El Trimestre Económico*, FCE, Oct. 1948, *XV*, 3, 396–412.

E. Padilla, "La Dinámica de la Economía Mexicana y el Equilibrio Monetario," *El Trimestre Económico*, July-Sept. 1958, *XXV*, 99, 349–77.

P. Padilla, "Elasticidad en las Ofertas de la Producción Agrícola Total en la República Mexicana," *Boletín de Estudios Especiales*, July 19, 1958, *XI*, 132, 297–309.

E. Perez Lopez, "Importancia para México del Turismo Extranjero," *Comercio Exterior*, Jan. 1961, 23–5.

———, "Analisis de la Autoridad y Estructura de la Proyectada Comisión Federal de Planeación," *Comercio Exterior*, Nov. 1963, 820–822.

I. Pichardo Pagasa, "Planeación del Desarrollo y la Reforma de la Administración Pública," *Comercio Exterior*, Aug. 1964, 533–39.

F. Rosenzweig Hernandez, "Programación y Desarrollo Agropecuario en México," *Comercio Exterior*, Oct. 1963, 726–28.

J. Saenz, "Problemas Monetarios," *Comercio Exterior*, Oct. 1957, 535–38.

———, "El Principio de Lucrecio (O Algunos Aspectos Monetarios y Fiscales del Desarrollo," *El Trimestre Económico*, FCE, Oct.-Dec. 1958, XXV, 100, 621–37.

W. Sedwitz, "La Devaluación: Su Génesis y Consecuenciase" *Revista Fiscal y Financiera*, April 1957, 11–34.

B. N. Siegel, *Inflación y Desarrollo: Las Experiencias de Mexico*, México 1960.

R. Stavenhagen, "Social Aspects of Agrarian Structure in México," in *The International Scene: Current Trends in the Social Sciences*, 1966, pp. 463–77.

R. Torres Gaitan, "Aspectos de la Política de Comercio Exterior Mexicana en la Décima Conferencia Interamericana," *Revista de Economía*, Vol. XVII, Num. 6, México, June 15, 1954, *XVII*, 6, 165–78.

———, "Panorama del Comercio Exterior," *Investigación Económica*, ENE, 1961, *XXI*, 84, 763–87.

———, "La Política de Comercio Exterior," *Comercio Exterior*, April 1962, 211–17.

V. L. Urquidi, "El Impuesto sobre la Renta en el Desarrollo Económico en México," *El Trimestre Económico*, FCE, Oct.-Dec. 1956, *XXIII*, 4, 424:37.

———, "Problemas Fundamentales de la Economía Mexicana," *Cuadernos Americanos*, Jan.-Feb. 1961, *CXIV*, 1, 69:103.

———, "México Ante los Mercados Comunes," *Comercio Exterior*, July 1962, 421–22.

G. R. Velasco, "Reflecciones sobre las Inversiones Extranjeras," *Revista Bancaria*, Jan.-Dec. 1955, 16–21.

R. Vernon, "México Ante los Mercados Comunes (Otro Punto de Vista)," *Comercio Exterior*, Aug. 1962, 527–529.

M. S. Wionczek, "Las Opiniones Extranjeras sobre la Devaluación," *Revista de Economía*, ENE, May 15, 1954, *XVII*, 5, 148–52.

———, "Los Bancos Centrales y los Acuerdos Regionales de Integración en América Latina," *Comercio Exterior*, Nov. 1963, 845–848.

———, "Integración Económica Regional: Los Factores no Económicos," *Comercio Exterior*, Sept. 1966a, 688–91.

———, "Apreciaciones sobre el Desastre de

Montevideo," *Comercio Exterior,* Dec. 1966b, 916–18.

———, "Nacionalismo Mexicano e Inversión Extranjera," *Comercio Exterior,* Dec. 1967, 980–85.

———, "La Transmisión de la Tecnología a los Países en Desarrollo: Proyecto de un Estudio sobre México," *Comercio Exterior,* May 1968, 404–13.

J. Zamora Batiz, "Comentarios sobre la Planificación Mexicana, *Revista de Economía,* ENE, Oct. 1967, *XXX,* 10, 313–14.

Birf y Nacional Financiera, *El Desarrollo Económico de México y su Capacidad para Absorber Capital del Exterior,* FCE, México 1953.

Revista de Economia, Editorial, May 15 1954, *XVII,* 5, 133–34.

United Nations Economic and Social Council, *External Disequilibrium in the Economic Development of Latin America: The Case of Mexico,* Vols. I and II, New York 1957.

United Nations, Economic Commission for Latin America, *Theoretical and Practical Problems of Economic Growth,* New York 1950.

Yugoslav Economic Policy in the Post–War Period: Problems, Ideas, Institutional Developments

By Branko Horvat

TABLE OF CONTENTS

Yugoslav Economic Policy in the Post–War Period: Problems, Ideas, Institutional Developments

By BRANKO HORVAT*

Introduction

Yugoslavia has been described as one country with two alphabets, three religions, four languages, five nations and six federal states called republics. One might add that the country has a population of twenty million and that it lies in the heart of the Balkans, with all that this connotes historically. For centuries the Balkans have been a meeting place of three world cultures and three powerful religions: the Catholic West, the Greek Orthodox East and the Moslem South. In terms of contemporary economic organization we may refer to the capitalist West, the centrally planned East and the undeveloped South.

All these influences have been felt. A rather turbulent life was to be expected in a country so located and having these characteristics. The present generation of Yugoslavs has experienced all three known modern economic systems: capitalism before the war, centrally planned economy after the war and self–government socialism in more recent years. The last–mentioned system is their own innovation and so far the only one of its kind in existence. The same generation has also experienced all four modern political regimes: bourgeois democracy (in the form of a constitutional monarchy and multi–party system) before the war, fascism during the war, 2 one–party state immediately after the war, and self–government democracy which is now in the process of being developed. It has also lived through a partisan national liberation war and a revolution. After the war a centralized kingdom was replaced by a federal republic, and in two decades the country had three constitutions. Finally, the same generation has experienced three different economic epochs: a pre–industrial stage before the war, rapid industrialization in the two decades after the war and the recently begun stage of a modern industrial economy approaching the Western European level. Before the war, 77 percent of the population were peasants and 40 percent were illiterate. A few economic indicators will suffice to indicate the economic development that has taken place since then: (see table 1)

Illiterates still constitute close to one fifth of the adult population, but at the same time with 11 university and college students per 1000 of population the country has moved close to the very top of the world list.

Such a tremendous pace of change virtually destroyed all traditions, but it also created a new one, a tradition of no tradi-

* Director, Institute of Economic Sciences, Belgrade.
I am grateful to Helen Kramer for linguistic assistance and helpful comments.

TABLE 1

	Before the war Yugoslavia	Before the war Western Europe[a]	1968 Yugoslavia
Production per capita:			
Electric energy, KWH	80	500–1300	1000
Crude steel, kg.	17	150–300	96
Cement, kg.	60	100–190	190
Cotton yarn, kg.	1.3	5–11	5
Energy kg.	180	2100–4300	1030
Fertilizers, kg.	3	20–65	96
Sugar, kg.	5	24–47	25
Stocks per 1000 of population:			
Radio sets	9	110–200	160
Automobiles	1	17–50	20

[a] France, Germany, Sweden, United Kingdom.
Sources: *SGS–1969. U.N. Statistical Yearbook 1956.*

tion, a tradition of change. In line with that the 1958 Program of the League of Communists—the heir of the Yugoslav Communist Party—ends with the words:

> Nothing that has been created must be so sacred for us that it cannot be surpassed and cede its place to what is still more progressive, more free, more human.

In such circumstances economic discussion displayed certain unusual features which make formal presentation somewhat difficult. Until about 1960 most of the discussion was either not put on paper, or at least not published. Further, professional articles made practically no use of references. There was a feeling of a complete break with the past, and so there was nothing to be referred to. In the same period professional literature was almost completely descriptive. That was due partly to the fact that the first university departments in economics were established only after the war. It is said that 90 percent of all scientists who have ever lived, live today. As far as Yugoslav economists are concerned, this percentage is virtually 100.

The second reason for the lack of analytical literature is to be found in the fact that there was hardly any time left for analysis. Economists were busy changing organization, institutions, and policies and keeping themselves informed about all these changes. Unless one had the inclinations of an economic historian, it did not make much sense to engage in a long–term research project. Before the book came off the press, the system had already been changed. Thus for quite some time professional economists were just describing what was happening. Description always precedes analysis.

Finally, until recently attention was mainly focussed on what Yugoslav economists call the "economic system." Economic policy in the traditional sense—the use of a set of instruments to achieve desired results in a *given* framework—hardly existed. Problems encountered were generally solved by changing the institutional framework itself. For a long time, and to a certain extent even today, economic policy consisted of an endless series of reorganizations. The search for an appropriate economic system was the main preoccupation of economic policy.

After 1960 economic organization began to assume a more permanent shape and economic discussion began to take a more familiar form. Since then use has been made of references in articles, ties with the

past and with the rest of the world have been established, economic debates have become frequent and lively, professional competence has increased, and a specifically Yugoslav theory of economic policy is now beginning to emerge.

I. *Three Economic Reforms*

Centrally Planned Economy

Institutional Development: Of all European countries occupied by the fascist invaders, Yugoslavia was the only one to liberate herself by her own forces. The National–Liberation War coincided with a genuine Social Revolution. This meant two things: an unbelievably high morale, the readiness to assault heavens—as a poet said—and also a hardly imaginable degree of devastation of the country. About 1.7 million people were killed in the battles, in concentration camps, by penal expeditions and by domestic quislings. One in every nine inhabitants disappeared in this way. Almost two fifths of the manufacturing industry was destroyed or seriously damaged. About three and a half out of fifteen million people were left without shelter. The loss of national wealth amounted to 17 percent of the total war damage suffered by eighteen countries represented at the Paris Reparations Conference in 1945 (Informativni priručnik, 1948, pp. 27–29). Apart from all this, the financial system of the country was in a chaotic state; divided and occupied by various aggressive neighbors, the country was left with seven kinds of currencies (German marks, Italian, liras, Hungarian pöngos, Bulgarian levas, Albanian francs, Serbian dinars and Croatian kunas).

The first task of the new government was to repair war damages as fast as possible and to organize the economy on what were considered to be socialist principles. For this purpose all available human and material resources were centralized, and with enormous efforts and great enthusiasm by 1947 the prewar output was achieved. The program of socialist reconstruction was carried out by means of legislative and political activities.

Yugoslavia was a peasant country. Peasants participated in the National Liberation War en mass. Agrarian reform, initiated already after the First World War, had never been fully implemented because of the opposition of the ruling classes. No wonder that one of the first moves of the new state was to undertake a radical agrarian reform. The land was to be given to those who tilled it. In less than three months after the end of the war a law was passed that took away the arable land in excess of 87 acres from farmers, in excess of 12 acres from nonfarmers. Big landowners lost their land without compensation. The land that was acquired in this way was distributed among poor peasants, who received about one half of the total land, to cooperatives and state farms (Dobrinčić *et al.*, 1951, pp. 53–54).

The next crucial move, undertaken in 1946, was nationalization of private capital in industry, mining, transport, banking and wholesale trade establishments. In 1948 nationalization was extended to retail trade and catering and in 1958 to houses with more than three apartments. About one half of the Yugoslav economy, outside agriculture, had been owned by foreign capital. Of the remainder, a sizable part had been owned by the Royal government which possessed coal and iron ore mines, forests and the largest agricultural estates; enjoyed a monopoly in retail trade of tobacco, salt, matches and kerosene; and was the largest wholesale trader, transporter, importer and exporter, banker, building entrepreneur and real estate owner (Bičanić, 1962a, p. 78). Since a number of private businessmen collaborated with the fascist invader and quisling governments, their property was confiscated. Those who took part in the

Resistance—and Communist Party members did that as a matter of course—very often gave away their property without asking for compensation. And, as was already noted, many business establishments were destroyed or damaged. In such circumstances complete nationalization was politically possible, was relatively easy to carry out and did not represent an excessive financial burden.

The next move was to introduce planning by a law in June, 1946. Plans were prepared by the Federal Planning Commission, responsible directly to the Federal Government.

Everything was now ready for the new Constitution which was adopted in 1946, and in which Article 15 read: "In order to protect the essential interests of the people, increase national welfare and make proper use of all economic potentials, the state directs economic life and development through a general economic plan relying on the state and cooperative sector and exercising general control over the private sector in the economy." This paragraph may be considered as both the definition and the inauguration of a specific socio-economic system, later to be known as administrative socialism or étatism.

The year 1947 brought the First Five Year Plan which was to lay the foundation for the future industrialized and developed Yugoslavia. The Plan was extremely ambitious—national income was to be doubled as compared with the pre-war level—but in the first eighteen months it was quite successfully carried out. It appeared as though the period of violent revolutionary upheavals was over and the country settled on a well defined and predictable course of economic and social development.

However, for Yugoslavia, history had always had some surprise in store. This time the surprise was more than unexpected: it was a complete shock. In the first half of 1948 Stalin accused Yugoslav Party leaders of revisionism and anti-sovietism. Yugoslavs rejected the accusation, and soon afterwards the Cominform countries launched a full scale political and economic attack. The Yugoslav Communist Party was excommunicated from the "family of brotherly parties," various treaties were abrogated unilaterally, development loans cancelled, trade with Yugoslavia amounting to about one half of her total foreign trade reduced to virtually nothing by the middle of 1949, and a complete economic boycott established.

The first reaction on the Yugoslav side was a somewhat naive but understandable attempt to prove that Stalin and others must have been misinformed, that no one questioned orthodoxy in organizing a socialist economy, that state ownership and central planning were keystones of the system. Motivated by consideration of this sort, in January 1949 the Central Committee of the Party decided to accelerate the collectivization of agriculture. Already in an income tax law, passed in August 1948, it was stated that "the rate of taxation should be such as to foster peasants' work cooperatives by means of lower taxes." A law on cooperatives, passed in June 1949, provided a legal framework for various types of cooperatives. Individual peasants were free not to join cooperatives if they chose. But by political propaganda and various administrative and financial devices, the authorities exerted strong pressure on them to join, and they did so in great numbers.

Meanwhile the organization of the economy was modeled after the Soviet pattern. The state budget absorbed the greater part of national income. The state apparatus was running the economy directly by means of ministries and directorates. By 1950 organizational development reached the stage at which the Yugoslav economy could be considered as

a model of an administratively run or centrally planned economy (Milić, 1951, pp. 126–70). This was also the climax. Already in 1950 a new development set in. The following year a complete overhauling of the economic system was in full swing. And by the end of 1951, the centrally planned economy belonged to history.

Discussion: The ideas and theories that served as guidelines in organizing the Yugoslav economy immediately after the war are to be sought in pre-war discussions among Yugoslav Marxists. They followed the well known orthodox viewpoint according to which socialism meant state ownership cum central planning. Immediately after the war there was so much to to that little time was left for leisurely reflections. Besides, everything seemed pretty clear, both theoretically and practically. One could rely on Marxist literature and on the experience of the Soviet Union, the first socialist country. What mattered most in those days was fast economic growth. And the Soviet Union showed how to achieve it.

But copying the Soviet blueprint did not produce quite the results expected. Furthermore, the savage attack of the Cominform countries forced people to reconsider their ideological positions quite thoroughly. And so preconditions were created for the emergence of a Yugoslav version of socialism.

Economic discussion before 1952 was dominated by two themes: planning for fast growth and the search for an authentic socialism. Since the former theme will be dealt with in the chapter on planning, we shall focus attention here on the latter.

The older theory maintained that in socialism there would be no market and no prices. After the Revolution, Yugoslavia was going through a period of transition between capitalism and socialism. In this period commodity relationships were still necessary because of the existence of private ownership and because labor was still heterogenous (Kidrič, 1949). Boris Kidrič—a statesman[1] who was to dominate the economic thinking of the country until his premature death in 1952—maintained that only state ownership was truly socialist (1950a, p. 8), that "the state sector was the highest form of our social ownership . . . " (1950a, p. 8). The same opinion is still held by most economists in the Soviet sphere of influence. In Yugoslavia it did not survive beyond 1950. Consistent with the above reasoning was the extolling of the significance of state planning. R. Uvalić (1948, p. 20), Kidrič (1949, p. 42), S. Kraigher (1950, p. 12) and others repeated the familiar thesis of Soviet economists about planning as a fundamental law of socialist economics. A few years later this theory was to be described as a voluntarist fallacy.

Rereading of Marx and Engles showed the possibility of great confusion in interpretation. Marx and Engels wrote seldom and very little about socialism. What they wrote amounted to two groups of statements: one dealing with the organizational form of a socialist economy, the other with the essential social characteristics of a socialist system. Marx and Engels maintained that commodity relations and the market would disappear along with private ownership; there would be comprehensive planning: production and distribution would be organized without the mediating role of money. For many decades it seemed obvious that comprehensive planning meant central planning exercised by the government, and that the absence of private ownership meant state ownership. In 1950 it was discovered that Marx had never drawn

[1] He soon became the President of the Economic Council of the Government and the Chairman of the Federal Planning Commission. He was also a member of the top Party leadership.

the last conclusion. In fact, and here they argued about essential characteristics of socialism, Marx and Engels denounced the state, argued that it would wither away in a classless society, talked about the self-government of producers, and asserted that "... a worker is free only when he becomes the owner of his means of production." Marx's insistance on the freedom of the individual was discovered in a statement, which was later entered into the Party program, and which reads: "The old bourgeois society with its classes and class antagonism is being replaced by an association in which development of every individual is a precondition for the free development of all" (Horvat, 1969a, pp. 105–17).

Far from being truly socialist, state ownership turned out to be a remnant of capitalism, characteristic of backward countries that are building socialism, and likely to generate dangerous bureaucratic deviations (Dobrinčić *et al.*, 1951, pp. 16–18). In 1950 Kidrič wrote: "State socialism represents ... only the first and the shortest step of Socialist Revolution ... Persisting in state (bureaucratic) socialism ... unavoidably leads to an increase and strengthening of privileged bureaucracy as a social parasite, to a suppression ... of socialist democracy and to a general degeneration of the system into ... state capitalism ... The building of socialism categorically requires the development of socialist democracy and a bold transformation of state socialism into a free association of direct producers" (Kidrič, 1950b, pp. 5–6).

Very soon a similar position was accepted by practically all Yugoslav social scientists. M. Novak wrote that to keep state ownership would mean" ... not the abolition of the proletariat but the transformation of all people into proletarians, not the abolition of capital but its general rule in which a specific exploitation can be and necessarily will be developed" (Novak, 1955, p. 92). Approaching the problem from a different point of view, N. Pašić came to the conclusion: "In the past state intervention in the economy was erroneously identified with socialism. If this criterion were applied to the last several decades, it would bring into socialist ranks all eminent capitalist politicians of recent times, from Baldwin and Roosevelt to Hitler and de Gaulle" (Pašić, 1957, p. 11). A. Dragičević wrote: "Nationalization of means of production and planning are preconditions of socialism, but *only preconditions* and nothing more. In order to achieve "fully developed" socialism, many more additional factors are required, in the first instance a socialist development of political relations and of economic structure of the society" (1957, p. 218). Similarly, P. Kovač and Dj. Miljević observed that "state ownership and state management by themselves lead to small or no change in the position of the producer in the production process and in his right to participate in the management of the economy.... In the countries in which socialist Revolution was victorious, the state, instead of becoming an organ of the working people, may and does become an organ of the state and party apparatus, which rules on behalf of the working people" (1958, p. 13). R. Milić observes that "state socialism in the USSR through bureaucratic socialism develops into state capitalism...." (1951, p. 21). These statements are not quite so novel as they might sound. Already half a century ago Z. Fabri, in connection with a book by Lenin, wrote: "If the state becomes an owner, we shall have state capitalism and not socialism.... Under state-ownership all proletarians would become workers hired by the state instead of by private capitalists. The state would be an exploiter and that means that an entire crowd of higher and lower managers and an entire

bureaucracy with all its hierarchical strata would create a new ruling and exploiting class. It looks as if something similar has already been happening in Russia . . . (Stanovčić and Stojanović, 1966, p. 164).

Lately there has been a tendency to replace "state capitalism" by an emotionally more neutral term "étatism" (Stanovčić and Stajanović, 1966, pp. 328–36; Pećujlić, 1967). The most radical in this respect is S. Stojanović, a philosopher by profession: "The term étatism denotes a system based on state ownership of means of production and state management of production and other social activities. The state apparatus represents a new ruling class. As a collective owner of means of production it employs and exploits labor. The personal share of the members of the ruling class in the distribution of the surplus value is proportional to their position in the state hierarchy . . ." (1967, p. 35).

If the state is an institution alien to socialism, who is to organize the economic process? Clearly, the only available alternative is that this task be undertaken by producers themselves. Centralization as the principle of organization is to be replaced by decentralization, centrally managed economy by a self-government economy. In the middle of 1950 a law was passed by which workers' councils were created. The draft of the law was introduced to the Federal Assembly in a speech by President Tito who said: "The slogan, the factories to the workers, the land to the peasants, is not any abstract propaganda slogan, but one which has deep meaning. It contains in itself the whole program of socialist relations in production and also in regard to social property and the rights and obligations of the workers, and therefore it can be and must be realized in practice, if we really desire to build socialism" quoted according to (Bilandžić, 1967, p. 69). By 1952 the new economic system was already in operation.

Decentralization

Institutional Development: The preparation for the New Economic System—as it was called—started with the Law on Management of Government Business Enterprises and Economic Associations by Workers' Collectives enacted in July 1950, and ended with the Constitutional Law on Principles of the Social and Political System of Yugoslavia, accepted by the Federal Assembly in 1953. The New Economic System (NES) became operational in 1952. It was transitional in character and lasted until 1960. During these eight years the country achieved the highest rate of growth in the world: per capita gross national product expanded at the rate of 8.5 percent per annum, agricultural output at the rate of 8.9 percent, industrial output at the rate of 13.4 percent (Horvat, 1963; Popov, 1968, pp. 363–64).

The law postulated that workers' collectives conduct all activities of their respective enterprises through their managing organs, Workers' Councils and Managing Boards. The Workers' Council was to be elected by all employees of an enterprise in a secret ballot. J. A. Schumpeter once remarked: "Wild socializations—a term that has acquired official standing—are attempts by workmen of each plant to supersede the management and to take matters into their own hands. These are the nightmare of every responsible socialist" (1950, p. 226). Such a nightmare was now made legal and obligatory by an act of the Belgrade National Assembly. "The principle of producers' self-management"— explains E. Kardelj, a social scientist and one of the most active political leaders—"is the starting point of every socialist politics. . . . Revolution that fails to open the door to such a development inevitably must . . . stagnate in state capitalist forms and in a bureaucratic despotism" (Kardelj *et al.*, 1956, p. 17).

In 1951 the government was busy dismantling the central planning apparatus with its ministries, directorates and administratively fixed prices. The last directorates disappeared in 1952. On December 30, 1951, a Law on Planned Management of the National Economy was passed. It replaced detailed central planning of production by planning of so-called basic proportions such as the rate of accumulation and the distribution of investment. Enterprises acquired a large degree of autonomy. In 1951 there existed numerous categories of market and planned prices. This was all replaced by a single price structure which with certain exceptions was to be regulated by the market. The rate of exchange was made more realistic by devaluing the dinar six times. And so in January, 1952, the economy was ready to embark upon a new road of decentralization.

Once it was recognized that the essential features of socialism consisted in individual freedom and the autonomy of self-governing collectives, two important consequences followed. First, the political monopoly of the state and party apparatus became incompatible with the so-conceived social system. Second, in order to be really autonomous, working collectives had to have full command over the economic factors determining their position. The former consideration led to a gradual transformation of the Communist Party from a classical political party into what I called an association of political activists (Horvat, 1969a, p. 261). The process was initiated in 1952 when the Sixth Congress of the Party changed its name to the League of Communists. The latter consideration led to a market economy with, it was intended, a minimum of government intervention.

In 1952 and 1953 several laws were passed regulating the formation, operation and termination of business enterprises. The enterprises could be set up even by a group of citizens. The director was to be appointed on a competetive basis by a joint commission of the Workers' Council and the local government. Unsuccesful enterprises could go bankrupt.

In agriculture the collectivization drive had increased the number of peasants' work cooperatives, but with its compulsory deliveries, administrative controls and the rest it depressed output.

Once the idea of an all-embracing administrative state control was abandoned, it was useless to insist on collectivization in agriculture, even more so because of the poor economic results. Ž. Vidaković gives the following explanation: ". . . the massive participation of peasants in the armed phase of the Revolution and in setting up the revolutionary political power contributed to the failure of étatist-bureaucratic socialization of agriculture, since the social-politically active peasantry did not submissively accept the administrative methods of collectivization" (1967, p. 42). In 1953 the Law on Reorganization of the Peasants' Work Cooperatives made it easy for peasants to leave cooperatives and most of them used this opportunity. Those who remained were often poor peasants

TABLE 2

	Index of output	Number of work cooperatives
1930–1939	100	—
1948	103	1217
1949	103	6238
1950	75	6913
1951	106	6804
1952	75	4225
1953	106	1165
1954	94	896
1955	116	688
1956	97	561
1957	140	507
1964	170	16

(SZS, *Jugoslavija 1945–1964*, 1965, pp. 99, 111)

and that meant that the remaining cooperatives would not be viable. In order to prevent this from happening and also to curb income polarization in the villages, two months later the government carried out a new agrarian reform which reduced the land maximum to 25 acres. Since before the war nearly nine tenths of all peasant farms were smaller than 25 acres anyway, the new reform did not meet with much opposition. But the harmful effects of former policy were not wiped out. In Yugoslavia there was a long tradition of agricultural cooperatives. Forced collectivization did a great deal to discredit cooperatives. Later the general agricultural cooperatives, which were administratively established and given a monopoly in village trade, also contributed to the discouragement of a genuine cooperative movement.

After all these changes the six-year-old étatist constitution became grossly inappropriate, while the time was not yet ripe for a brand new constitution. The problem was solved by a Constitutional Law, passed in 1953. Its article four States: "Social ownership of means of production, the self-government of producers in the economy and the self-government of working people in the Commune, City and District represent the basis of the social and political system of the country. . . ."

As a consequence of the self-government principle, another very important innovation found its place in the Constitutional Law. It became known as the principle of the fusion of the political and economic sovereignty of the working people. The principle was implemented by creating the Council of Producers as a new house in the Assembly. The Council was composed of representatives of collectives of business enterprises.

In the following years the government was engaged primarily in perfecting the monetary and fiscal systems. Interest rates were applied and there was some experimentation with investment auctions. Commercial banks were added to the hitherto all embracing National Bank. Reserve requirements were introduced. Local governments acquired financial autonomy.

The First Five-Year Plan (1947–1951) was extended for a year, but never really completed. The period 1952–1956 was left with only annual plans. After NES was well established, the Second Five-Year Plan covering the period 1957–1961 was launched. It was carried out in less than four years.

Discussion: While the preceding period was mostly characterized by discussion of what was *not* socialism, the theoretical approach becomes more positive now. The discussion started by an exchange of opinions on the so-called Transition Period and ended with an analysis of what was to be known as self-government or associationist socialism.

Marx wrote that the revolutionary transformation of a capitalist into a communist society could not be carried out at once. Between the two socio-economic systems there must be a short transitional period, and the state of this period would be organized as a Dictatorship of the Proletariat. Marx's analysis looked plausible and in fact proved to be a good anticipation of what happened in Yugoslavia in the first two decades after the war (Horvat, 1969a, ch. II). Around 1952, and intermittently later, the main issue of the debate was whether socialism (considered to be the first of the two stages of a communist society) is to be included in or excluded from the Transition Period (Horvat, 1951; Novak, 1952 and 1955; Perović, 1953; Sirotković, 1951; Kiorač, 1951). The debate was highly scholastic, and yet the issue was of enormous practical importance. If Dictatorship of the Proletariat is interpreted as a form of political regime, and not as the class content of the govern-

ment (which is what Marx had in mind), the identification of socialism with the Transition Period will produce a command society. If the political regime is democratic, but the Transition Period extended to include socialism, the development of a classless society may be endlessly delayed. The issue was resolved in an indirect way after the essential characteristics of a self-government socialism had been elaborated.

Contrasting the Old (administrative) and New (self-government) Economic System, R. Bičanić[3] summarizes the actual developments by enumerating differences in goals, agents and means (1962a, pp. 44–47). The *goals* of the Old System were to achieve socialism by means of state power, to equalize the position of workers in relation to the state-owned means of production, and to achieve the new social order for its own sake. Individual interests of producers and consumers were subordinate to impersonal and superhuman goals of the economic system, and the state apparatus, entrusted with the achievement of this goal, was in a position to exploit the population. The New System presupposes the withering away of the state and the management of socialized property by workers, and makes the personal happiness of every individual a supreme goal.

As far as the agents are concerned, in the Old System there was centralized state management by means of a hierarchically organized state apparatus. The directives were passed down the line in an authoritarian way with little or no independence of enterprises. In the New System the state apparatus cannot interfere with the business of individual enterprises, which became autonomous. Decentralization was applied not only to economic, but also to social and political life. Authoritarianism was replaced by self-government as a basic principle of economic and social organization.

The means of the two systems are contrasted by Bičanić in the following way: state ownership vs. social ownership; central planning vs. social planning; administrative allocation of goods vs. market; administrative rules vs. financial instruments; administratively fixed wages vs. free disposition of the income of the working collectives; all-embracing state budget vs. the budget of the state administration decentralized and separated from the economic operations; consumption as a residual vs. consumption as an independent factor of development; collectivization vs. business cooperation of peasants and large agricultural estates.

In the period under consideration economists began to study intensively writings in the economics of socialism, particularly those of Western authors. This literature had hitherto been virtually unknown. I. Maksimović (1958), F. Černe (1960) and B. Horvat (1964) produced extensive critical accounts of earlier economic literature. Černe attempted to provide an acceptable definition of socialism. In his view socialism is characterized by the following three elements: (1) Equal rights of members of the community as producers. This implies social ownership. Element (1) is a precondition for (2) equal rights in terms of income distribution. This in turn implies distribution according to work. Both (1) and (2) are indispensable for the realization of (3) equal rights in political life. As citizens members of the community must enjoy political—Černe talks of socialist—democracy (1960, p. 281). It appears that socialism is essentially a philosophy of egalitarianism. Černe's definition, although never explicitly quoted —references are not popular in Yugoslavia —may be considered as commanding wide

[3] Bičanić completed his study early in 1961. Essentially the same comparative analysis had already been presented by M. Popović in 1952 (1952). Evidently, the system was being developed in a consistent way.

agreement among economists and other social scientists.

On a less abstract level, in an important article in 1953, Uvalić described the main intentions of NES (1954). In the administrative period output was expanded regardless of cost. Now fast growth was to be maintained but cost considerations had to play an important role in the determination of the structure of output. The law of value, i.e. the market, was to take care of that. But the operation of the law of value must be restricted in two important respects: income distribution and capital formation must be controlled. Otherwise, Uvalić warned, exploitation and market anarchy will reappear. These ideas were to dominate economic policy in the next decade. But clumsy bureaucratic and often incompetent controls of income distribution and capital formation were to become more and more irksome and irritating.

The relation between market and planning has become a recurrent theme in economic discussion. Usually market and planning are visualized as two different mechanisms. In the opinion of Černe the planning mechanism is to be used for long-run and general decisions, while short-run and partial decisions may be left to the market mechanism (1960, p. 11). A similar position was taken by J. Lavrač (1958). B. Jelić explores in more detail institutional arrangements necessary to harmonize the market and planning. He argues that unbalanced growth sometimes requires interventions even outside the general framework provided by the plan (1958).

By the end of the period (1958) under consideration, NES got its first theoretical rationalization in a book by the present author (Horvat, 1964). Since the socio-economic system is conceived as an association of business, political, etc., associations, I suggested that it be called Associationist Socialism. I pointed out that the old alleged incompatibility of market and planning was nothing more than an ideological fallacy. The market is just one and at that a very efficient—device of social planning. The integration of market and planning, social ownership and business autonomy of enterprises, produces a system with interesting new practical as well as theoretical features, First reactions towards this book were negative (Dragičevic, Stampar & Horvat, 1962; 1963). Insisting on consumer sovereignty was considered to represent the (negative) influence of Western welfare economics. Insisting on rigorous technical analysis was considered devoid of social content and so anti-Marxian. Insisting on market economy was considered to reflect the influence of the Western theory of free competition. The analysis of price formation, in which interest and rent played a certain well-defined role, was said to represent a bourgeois theory.

A similar critique was voiced by some socialist economists abroad. E. Mandel maintained that "there is a definite incompatibility between socialism—or, put otherwise, a classless society and a high degree of social equality and economic efficiency —and commodity production" (1967). This is so because commodity production inevitably generates social inequality and produces waste of economic resources. The reader was not told why this should be inevitable.

In this debate B. Ward came perhaps nearest to the truth. As to the method of analysis she says: "In value theory Horvat manages to produce more or less Marxian results from more or less neoclassical assumptions" (1967, p. 519). As to the substance of the theory she concludes: "Naturally enough this regime is essentially socialist; not surprisingly, it bears a more than casual resemblance to Yugoslavia. What is surprising is that it carries a more than expected measure of plausibility . . ."

(1967, p. 509). Most of the ideas developed in this 1958 book have by now been absorbed and seem self-evident. The latest reform is based on the market mechanism and the welfare of individuals as the main guiding principles.

Self-government Socialism

Institutional Development: The last phase in Yugoslav post-war socio-economic development was prepared by a series of political, economic and constitutional reforms in the period 1958–1963. This turbulent period was inaugurated by the new *Program of the League of Communists* in 1958. Here socialism is defined as: ". . . the social system based on socialized means of production in which social production is managed by associated direct producers, in which income is distributed according to the principle to each according to his work and in which, under the rule of the working class, itself being changed as a class, all social relations are gradually liberated from class antagonisms and all elements of exploitation of man by man" (*Program SkJ* 2a, 1958, p. 133). Thus the Yugoslav variant of socialism appears to imply social ownership, self-management in the economy, the absence of non-labor income and of exploitation. The term "working class," as explained a few years later by Kardelj, was to mean "all working people who are participating in the social process of labor and in socialist economic relations" (Kardelj, 1962, p. 1531).

By 1960 the second Five-Year Plan was successfully completed. The economy was booming, self-management in enterprises was already well established and the Program paved the way to an accelerated pace of changes. The new Five-Year Plan was prepared. The Society felt ready for a new important step forward. In 1961 three radical reforms were carried out. In order to increase the efficiency of the market organization and to improve the quality of goods produced, the hitherto virtually closed economy was to be made more susceptible to the influences of the world market. To achieve that, the system of multiple exchange rates was replaced by a customs tariff, the dinar was devalued, foreign trade was liberalized to a certain extent and the country became an associated member of GATT. Since developments in the field of money and banking were lagging behind the general institutional changes, an overhaul of the entire financial organization was undertaken. And finally, it seemed inappropriate for trade unions to continue to supervise wage levels and wage differentials in self-managed enterprises. And so this control was discontinued. Since then in this field, market competition has gone further than in any other modern economy. These three reforms inaugurated in 1961 the beginning of the third distinct phase of economic development.

By that time the country was institutionally ready for the new constitution which was promulgated in 1963. Explaining the aims of the constitution. Kardelj, one of its chief architects, said that it was "not only the constitution of the state but also a specific social charter which will provide the material basis, political framework and encouragement for the faster internal development of the system of social self-government and direct democracy" (1962, p. 1533). Self-management was extended to cover not only business but also non-profit organizations. It was generalized as a principle of self-government to be applied in all spheres of economic, social and political life. In order to achieve this, the Constitution invented a new institution: the work organization (rādna organizacija). Whenever people associate in order to work for a living, they create a work organization and represent a work union (radna zajednica) which en-

joys basic self-government rights constitutionally guaranteed. Work organizations include enterprises and other business establishments as well as educational, cultural, medical, social insurance and other public service establishments. As a consequence the "fusion principle" of the 1953 Constitutional Law was extended to cover all work unions, and the Assembly got three houses of work unions: for the economy, for education and culture and for health and social welfare.

The three reforms of 1961 were poorly prepared, partly inconsistent and badly implemented. As one might have expected, the sensitive market economy reacted violently. Everything went wrong: in one year the rate of growth of industrial output was reduced to one half of its 1960 level, imports soared, exports stagnated, wages went far ahead of productivity. The reformers, accustomed to a tardy half-administrative economy, were taken by surprise. Planners increased targets for 1962 in order to catch up with the Five-Year Plan goals—and were, of course, deeply disappointed. The recession was deepened. It became clear that the Plan would have to be abandoned. Administrators and political bodies were deeply disturbed. Conservative politicians and economists were busy explaining the failure of the market system and demanded that central planning be reintroduced.

Heavy pumping of money into the economy helped to generate recovery in the second half of 1962. In the next year the economy was back to its normal path of fast growth. The upswing continued into 1964 ending in a boom with heavy inflation and a great balance of payments deficit. The new recession brought a new reform. Throughout 1964 assemblies were busy discussing the principles of the new reform (Savezna Skupština, 1964). In the beginning of 1965 the government administration was set to work. By May, technical preparations were completed and in July the Federal Assembly enacted the package of laws inaugurating the reform (Savezna Skupština, 1965). Significantly enough, the solution of economic troubles was sought in further decentralization, perfection of self-government autonomy, development of a more competitive market and an integration into the world economy. What followed appeared to be a second, more radical and more consistent, edition of the 1961 reform. The reform started as an economic one, but very soon produced important social and political consequences. Multicentric planning could not help but produce a pluralistic society. Reform was in its essence a new stage of the revolution; so asserted V. Bakarić, president of the Croatian League of Communists (1967, p. 231). Self-government autonomy became firmly rooted in the Socialist Establishment.

Discussion: The reform of 1961—called also NES (II)—marked the beginning of a real academic discussion of economic matters. Up to that time institutional changes had been foo fast, and economists too few, so rigorous analysis and discussion had been replaced by descriptions.

The discussion started with an exchange between Uvalić and Bićanić. Uvalić reiterated his views that income distribution and capital formation could not be left to regulation by the market. So far, distribution according to work had encountered serious difficulties. The capital market, as a device for capital formation and allocation, was unacceptable because it would lead to group ownership. Social profitability and individual profitability were two different things. The individual interest of a collective was inferior and had to defer to the social interest (Uvalić, 1962). Bićanić objected that Uvalić did not distinguish clearly between what is commonly called the economic system, and the plan. The economic system (general

conditions for business conduct) used to be an instrument of the plan; now the relation had been reversed. (In fact, two years later a party congress would request explicitly that the plan become an instrument of the system instead of the system being accommodated to the plan (Šefer, 1968b, p. 29). Uvalić offered no guidance as to how to replace labor and capital markets. He really implied central planning, with operational freedom being left to planners and politicians and discipline being reserved for the rest. Bičanić feels that this is unacceptable. A modern economy is essentially polycentric and not monocentric (1963a).

In December 1962 the Association of Yugoslav Economists organized a debate in Belgrade about the draft of the new constitution (Ehonomist, 1962). A number of participants—R. Davidović, M. Macura, N. Čobeljić, K. Mihajlović—argued that the role of planning was underestimated in the draft constitution. Macura explained that this was so because economic problems were approached from the point of view of an enterprise, even an individual, instead of from the point of view of the economy as a whole (*Ekonomist,* 1962, p. 462). Čobeljić thought that planned market economy would in future be replaced by market planned economy (*Ekonomist,* 1962, p. 473). Mihajlović argued that, while consumer and intermediate goods markets worked well, investment goods and capital markets were notoriously imperfect and needed strict control (*Ekonomist,* 1962, p. 500). The debate reached its climax at another meeting a month later in Zagreb.

Futher discussion was prompted by the failure of the reform. The economy sank deeply into depression (relative to the standard Yugoslav state of affairs). From the beginning of 1961 to the middle of 1962 the annual rate of growth of industrial output dropped from 12 to 4 percent. The government was alarmed and asked a group of academic economists associated with a research institute to find out what had happened. This move set a precedent in the governmental attitude towards managing economic affairs. In a few months the group produced a report, popularly called *The Yellow Book* (Horvat, 1962a). Then the second, even more important, precedent was established: the government accepted the report.

The findings of the *Yellow Book* may be summarized as follows. Inefficient planning resulted in economic instability. The structure of supply failed to match the structure of demand, there was a downward shift in long-run export trends, there was a serious lack of skilled labor force. The inherently unstable economy was exposed to the simultaneous shocks of the three poorly prepared and badly implemented partial reforms cited above. The insistence on financial discipline created a serious shortage of money with strong deflationary effects. The abolition of income control led to wild increases of wages unrelated to productivity increases. The liberalization of foreign trade emphasized the fundamental importance of economic research as a basis for economic policy and the stability of the legal and policy framework as a precondition for efficient operations of enterprises in a market setting.

In the meantime, another research institution produced an analysis of the defects of the economic system. The report became known as *The White Book* and it criticized deficient planning, an imperfect market, arbitrariness in income distribution and inconsistencies in investment decisions (Dabčević *et al.*, 1962). Both documents were discussed in a meeting jointly organized by the Association of Economists and the Federal Planning Bureau in Zagreb in January 1963 (Savjetovanje, 1963). The former planning officials and a certain number of economists with a more

centralist orientation criticized the two documents. They questioned the possibility of efficient investment and a high rate of growth in a decentralized setting. They thought that the market necessarily led to a destruction of the socialist principle of of income distribution. Some of them pointed out that the classical conflict between the essentially social character of production and atomized decision-making lay at the bottom of all economic difficulties (Savjetovanje, 1963, p. 192). However, the majority of economists agreed on the necessity of further decentralization and the perfection of self-government autonomy. Since the Zagreb debate the basic principles of the development of the economic system have never been seriously questioned among Yugoslav economists.

The well know saying about doctors—the operation was successful but the patient died—might have been applied to discussions among Yugoslav economists: the causes of economic troubles had been well explained, but the reform was dead. It soon became clear that the entire experiment had to be repeated. And so it was, in 1965. The situation was rather complicated. "The casual observer is often puzzled," commented R. Bičanić. "Only a few years ago Yugoslavia was presented as an example of a country with one of the highest growth rates in the world, now the foremost aim of economic policy is to reduce investment. For more than a decade the socialist economy struggled against bureaucratic command; now an administrative price freeze has had to be introduced. It was the first country in the world to initiate workers management in business enterprises and to abolish the wage system; now there is discussion about whether this means too much or too little democracy. . . National problems were said to have been solved; and now the country is pregnant with increased tensions among the constituent nations, tensions newly created and socialist in origin. Efforts to find solutions to all these problems are now concentrated into two words: *The Reform.*" (Bičanić, 1966, pp. 633–34).

Bičanić and Dzeba (Dzeba and Belsać, 1965) the following aims of the reforms. The immediate purpose was to combat an increasing pace of inflation; to remove the chronic deficit in the balance of payments; to reduce all sorts of subsidies (for exports, unprofitable production, etc.) drastically in order to avoid the necessity of central administrative interventions; to correct price disparities in order to establish more efficient market relations and eliminate administrative controls. These were preconditions for some longer-term measures of structural change in the economy such as: revision of growth and investment policies; putting the productivity of the economy on an internationally competitive level; liberalization of foreign trade and elimination of the balance of payments deficit; convertibility of the currency in order to open the economy and expose it to the stimulating influences of the world market. In its broader social aspects, the reform was expected to impart a de-politicization of economic decisions; double the share of enterprises in the control of national income, reducing thereby the economic power of the state; to link the level of living to that of productivity; to increase the rationality of economic decision-making. Bičanić concludes that the fundamental aim was in fact "to build a model of a socialist system for a developed country, one which will be able to stand the competition of other developed countries without the constant tutelage of government machinery" (1966, p. 643). He and Horvat (Dobrinčić *et al.*, 1951) pointed out that this model is very different from the mixed economy of the welfare state.

The aim described was to be achieved by a process which Bičanić called the four

D's: Decentralization, De-étatization, Depolitization and Democratization.

As often happens, the ideas were good but the implementation was poor. The reform was politically much better prepared than the one in 1961, but not so economically. Economically it was based on a rather naive idea of the viability of the *laisser-faire* principle. Monetary policy appeared to be practically the only available device of economic policy. In order to stabilize prices, the government applied a credit squeeze. It worked, but it also produced deflation with unemployment and stagnation. From the beginning of 1965 to the middle of 1967 the annual rate of growth of industrial output dropped from twelve percent to minus one percent. Negative growth rates had not been known since 1952. The government thought that this was unavoidable, and that the reform "in its strategic aspects" proceeded as planned. Some economists and many businessmen were alarmed. For them, developments were catastrophic and certain to produce another failure. Soon economists were to discover the existence of business cycles. Since cycles had not been known to the government—it was held as self-evident that cycles could not exist in a socialist economy—the government proceeded to frame economic policy as if the cycles did not exist. The results of such an economic policy could not be encouraging.

The discovery of cycles proceeded in stages. The successive retardations of growth, described already in the *Yellow Book*, indicated that the Yugoslav economy might have been subject to cyclical fluctuations. The research undertaken in the Institute of Economic Studies confirmed the hypothesis. (This will be discussed more fully in Chapter 5.) In the Spring of 1967, in Ljubljana the Association of Economists held a meeting dedicated to problems of stabilization (Savjetovanje, 1966). Four papers dealt explicitly with business cycles. The research institute mentioned ventured to make a forecast of the lower turning point (1967), boom (1969) and recession (1970) of the current cycle which proved to be correct up to the time these lines were written (second half of 1969).

A couple of months after the Ljubljana meeting a public debate took place. It was focussed on the theme: "Economic Science and the Economy" (Institut, 1968a). Seven economists participated. A. Bajt raised the question of the responsibility for the reform and criticized the naive view that investment generated inflation. Z. Baletić evaluated the contention that there was a conflict between politicians and economists. Ž. Mrkušić analyzed the foreign trade equilibrium. Horvat pointed out a number of mistakes contained in currently popular economic reasoning (and, consequently, in economic policy), and in a separate article, which caused a newspaper explosion of discontent, calculated the losses due to cyclical instability. The output lost appeared to amount to about forty percent of the social product. The three remaining economists supported the official view that everything was more or less all right.

In February 1968 the Institute of Economic Studies organized an all-Yugoslav conference on the current economic situation. The study prepared for this occasion (Institute 1968b) described the cyclical mechanism operating in the Yugoslav economy and made a coherent proposal for an anti-cyclical policy. This was an important step forward. The proposal insisted on a combination of monetary and fiscal policies (the latter was virtually non-existent at that time); on a combination of price and income controls; and on the importance of the interrelations between aggregate demand and investment.

By the end of the same year another feature of the unsuccessful 1961 reform was re-

peated: two research institutes were officially asked to assess the implementation of the reform. There was, however, an interesting difference: this demand did not come from the government but from the Central Committee of the League of Communists. Two reports were prepared: the findings were more or less the same. I quote from the report that was published (Institut, 1969). This report found that in spite of a strong deflationary policy, prices were no more stable than they were before the reform; that the Five-Year Plan was not likely to be fulfilled; that the administrative control of prices was extended over a greater percentage of output than before the reform; that the liberalization trends in foreign trade were checked and reversed; that the balance of payments deficit was expanding; that the rate of saving was decreasing; that the losses and indebtedness of firms were increasing; that the rise in labor productivity was slightly retarded; and that unemployment was increasing beyond anything known in the country in the past two decades. Elaborating its early prognosis in more detail, the Institute predicted an acceleration of growth in the first half of 1969 (to a rate some sixty percent higher than the one forecast by the Federal Planning Bureau), an inflationary pressure in the second half and the downturn of the cycle and the beginning of a new recession by the end of 1969 or in the first half of 1970. The first two forecasts proved to be correct, the last had still the status of forecast at the time these lines were written. A few months later V. Rajković undertook to analyze the unpublished papers prepared by the administration as a basis for the reform. Rajković came to the conclusion that none of the important goals had been achieved in a satisfactory way (1969/70, p. 47).

Once again the ominous question was posed: What had happened? A careful analysis of developments seems to suggest the following answer. Economic growth and institutional changes were too rapid for the government apparatus and other organs of economic policy to be able to cope with efficiently. Almost overnight a backward Balkan country reached a European standard of economic development, and an adminstrative economy was transformed into a market economy. At the same time responsible authorities often lacked the necessary understanding of how a modern market economy operated. If to all that we add the pioneering in the system of self-government—nonexistent anywhere else in the world—it becomes clear that the complexities of the socio-economic environment have increased enormously and that it will take some time before the organizational framework is adapted, the necessary knowledge is accumulated and the new social system begins to operate smoothly (Institut, 1968a, 1969; Horvat, 1968a).

II. *Planning*

Four Five-Year Plans

The rationale for central planning was explained in Chapter 1. By 1947 the machinery for central planning was completed. Hierarchically organized planning commissions on various levels—federal, state, district and city—were entrusted with comprehensive planning in their respective territories. The operational planning and implementation was carried out by ministries and then down the line by general and chief directorates, and planning sections in the enterprises. Annual plans were broken down into quarterly, monthly and ten-day plans. In 1949 about 13,000 groups of commodities were planned (Čalic, 1948, p. 15). In the same year the state budget comprised two thirds of the national income (Kidrič, 1960, p. 453). Every enterprise had to send to the su-

perior authorities 600–800 different reports per year. The annual economic plan weighed some 3,300 pounds (Bičanić, 1957, p. 65). Supplies and customers were assigned to every enterprise in advance. Since these administrative allocations were not quite perfect, the enterprises were asked "to find their ways." The planning authorities would provide them more money then they wanted and would ask them to spend it. As prices were fixed, spending money meant finding raw materials and investment goods necessary for the fulfilment of the plan. In a market economy one endeavors to save money, in the centrally planned economy one is at great pains to spend it: in the former selling is the most difficult task, in the latter buying is the greatest worry of businessmen.

The economy was run as one single mammoth enterprise. That required establishing a system of continuous control of operations of all enterprises. In 1948 Kidrič voiced complaints against those who considered that there was no need for daily reporting and who were satisfied with ten-day reporting (Kidrič, 1960, p. 468). A number of years later J. Stanovnik, now Secretary of the U. N. Economic Commission for Europe, at a lecture delivered to Swedish economists in Stockholm, was asked what sort of devices were used to implement plans in Yugoslavia. He answered: "Telephones!."

The first Five-Year Plan covered the period from 1947 through 1951. It proclaimed four main goals:

(1) to overcome economic and technological backwardness;
(2) to strengthen the economic and military power of the country;
(3) to strengthen and develop the socialist sector of the economy;
(4) to increase the general welfare of the population.

Consumption was taken care of, but it was last in the order of priorities. The goals enumerated were to be achieved by an explosive increase of output; compared with the pre-war level, national income was to increase 1.9 times, agricultural output 1.5 times, industrial output 4.9 times. However, due to poor statistics, the pre-war level must have been greatly underestimated and the three targets were achieved only by 1954, 1959 and 1961 respectively.

At first the implementation of the plan proceeded in a satisfactory way, though not as well as was generally believed.[3] In 1949 the economic blockade of the Cominform countries forced Yugoslavia to search for trading outlets for about one half of her exports and to secure the same proportion of imports from other sources. Although substantial foreign aid was secured two years later, this sudden reorientation of foreign trade had stifling effects on growth. The next blow came from nature; in 1960 a severe drought reduced agricultural output by one third. Collectivization also helped to aggravate agricultural problems. The radical economic reorganization in 1951 could only complicate matters. Industrial output fell by four percent in 1951, and by one more percent in 1952. The plan was extended for a year, but that was already pointless, and the report on

[3] Thus V. Begović reports about the overfulfillment of the first half of the Five-Year Plan (1949). But later statistical estimates showed that the data produced by the Federal Planning Commission (Informativni Pritučnik pp. 251, 484) were inflated. Thus for the output of manufacturing, mining and power plants the differences are as follows:

	Indices		
	1948 / 1946	1949 / 1948	1950 / 1949
Federal Planning Commission	267	116.6	106.3
Federal Statistical Office (later estimates)	190	111	103

the fulfilment of the First Five-Year Plan was never published.

And yet, if not a full success, the Plan was far from being a failure. It generated output substantially above the pre-war level, it raised the share of gross investment in fixed assets to 33 percent of gross national product (material product definition; close to 30 percent on the SNA definition) and created entire new industries.

In 1952 rigid central planning was replaced by "planning by global proportions." These proportions were: minimum use of output capacity and the corresponding wage fund, profits as a percentage of the wage bill (a device for wage planning), basic capital formation, taxes and allocation of budgetary resources (Vučković, 1952, p. 31). In this way, the central plan was expected to regulate general economic activity without administrative orders, by influencing the rate of growth and the proportion between investment and consumption, and by effecting structural changes in the economy (Jelić, 1961). The old Planning Commission—which acted as a superministry controlling the activities of all economic ministries and was in charge of the overall implementation of the plan (Djordjević, 1965)—was replaced by the Federal Planning Bureau, an expert institution with no administrative powers. Republics, districts (later communes) and enterprises would produce their plans independently. State planning became social planning which meant wide consultations among all interested parties, inclusion of non-profit institutions and independence of enterprise plans.

The next three years were used to complete the key investment projects of the Five-Year Plan in annual installments. In the discussion about the 1955 Plan the new mood was already apparent; agriculture looked neglected, investment too large and onesided (Popović, 1964, pp. 147, 150). By the end of that year M. Popović could say in the Federal Assembly that one period of economic development was completed (1964, p. 160). The year 1956 was used to prepare the Second Five-Year Plan for the period 1957-1961. In this plan increase of consumption already ranks third among the five main goals (Lovrenović, 1963, p. 220). Growth of investment was somewhat retarded and its structure radically changed. The share of industrial investment was substantially reduced in order to double the share of agriculture and increase the shares of transport and trade (Popović, 1964, p. 211). Within manufacturing, consumer goods industries were to expand faster. So-called non-productive investment in social overhead capital was also accelerated. All these changes proved beneficial and the plan was carried out in four years. The planning system seemed to be well adapted to the needs of the economy and worked satisfactorily. This system was described by J. Sirotković (1961), S. Dabčević (1963), and Jelić (1962).

The first plan distorted the structure of the economy by emphasizing capital formation in heavy industries. The second one undertook to make corrections but went to the other extreme by overexpanding consumer goods industries. Thus, the third plan was left with the task of redressing the balance again by accelerating investment in power generation, metallurgy and intermediate goods industries. These fluctuations in investment induced Čobeljić and R. Stojanović to invent a theory of investment cycles inherent in a socialist economy with an uneven pace of technological progress (1966). Z. Baletić, Bajt (1969) and others criticized this theory as unacceptable since mistakes in planning are attributable to ignorance and not necessarily to socialism, and that technological progress is rather innocent in this respect.

The Third Five-Year Plan for the period

1961 through 1965 endeavored to accelerate the growth of output even further. Personal consumption ranked second among the goals (Lovrenović, 1963, p. 221). The Plan was hardly launched when the country found itself in the middle of a recession, the reasons for which were explained in the previous chapter. The Plan was doomed to fail. In order to avoid unpleasant discussions, the Federal Assembly decided to replace it by a seven-year plan covering the period 1964–1970. For that purpose the Assembly passed a Resolution in which the basic political and economic goals of the new plan were defined as follows (*Yugoslav Survey*, 1964):

(1) steady rise of the level of living, in the first place of personal consumption, and higher share of personal incomes in national income;
(2) catching up with international standards of productive efficiency and labor productivity;
(3) expansion of external trade through more intensive inclusion of Yugoslavia in the international division of labor;
(4) accelerated development of underdeveloped areas;
(5) further development of socialist soviety by strengthening the role of direct producers and working organizations in the management of productive forces.

A comparison of these goals with those of the First Five-Year Plan shows very clearly the distance that separates social planning from state planning. The welfare of individuals is moved from the end to the beginning of the priority list.[4] Behind this change one finds the philosophy which holds that economic welfare is both the purpose and the most powerful incentive for production. An autarchical orientation is replaced by openness towards the world market and international influences. The measure of the perfection of a socialist economy is no more to be found in increasing the share of the state in the national capital but in the development of self-government. Yet, the First Plan and the Resolution had one thing in common: neither of them was implemented.

The Resolution in fact foreshadowed the Reform of 1965. The changes in economic institutions were so radical that it became necessary to prepare a new Five-Year Plan for the period 1966–1970. The Plan incorporated the goals of the Resolution. It envisaged a somewhat lower rate of growth of GNP (7.5–8.5 percent per year), a relatively modest expansion of manufacturing (9–10 percent), but a high rate of productivity increase (6–7 percent a year). Current analysis of the Federal Planning Bureau indicates that these targets are not likely to be achieved (Medenica, 1968).

Growth and Cycles

In order to be able to evaluate successes and failures in planning—and in economic policy in general—one has to have a look at some data. The following table summarizes the developments in terms of rates of growth of the most important statistical aggregates.

In the central planning period collectivization caused stagnation in agriculture and the economic boycott of the Cominform countries caused stagnation in exports. As a result total output grew slowly. In the second period the unfettered economy was in full swing in all spheres with an acceleration of growth in the second half of the period. Foreign trade expanded faster than output and exports faster than imports. In the third period agricultural out-

[4] Personal consumption was reduced at a rate of 2 percent annually in the period 1948–1952; it began to expand at [illegible] percent per annum in 1953–1956; it expanded at approximately the same rate as national income, at [illegible] percent, in 1957–1963; and its rate of growth surpassed that of national income afterwards (Šefer, 1965, pp. 207–209).

TABLE 3.—GROWTH OF THE YUGOSLAV ECONOMY 1946–1968
(RATES OF GROWTH, PERCENT PER ANNUM)

	Central Planning 1946–1952	Decentralization 1952–1960	Self-government 1960–1968
Gross National Product	2.3[a]	9.8	6.8[c]
Industrial output	12.9	13.4	7.9
Agricultural output	−3.1[a]	8.9	2.1
Export of Commodities	−3.1[b]	11.7	7.0
Import of Commodities	3.6[b]	9.7	7.0
Employment[d]	8.3[a]	6.9	2.4

[a] 1947–1952
[b] 1948–1952
[c] 1960–1967
[d] Persons employed outside private agriculture.
Sources: *Statistical Yearbooks.*

put caught up with domestic demand, while the European export markets became increasingly difficult to penetrate. The slowly expanding market for agricultural products reduced the rate of growth of agricultural output. Increased economic instability depressed the average rate of growth of manufacturing. As a result, the overall pace of expansion was reduced. In all these developments institutional factors, described before, played an important role. If one wants to judge the performance of the economy on the basis of a somewhat longer period, the period 1952–1968 appears to be the appropriate one. In these sixteen years, total output expanded three and one half times, manufacturing five times, agriculture two and a half times, foreign trade in commodities four times and employment outside private agriculture three times.

Since Yugoslavia has been so far the only country that has lived through three different economic systems—capitalist, étatist and self-government—in a relatively short period of time, it may be possible to evaluate the comparative efficiency of the three systems. Something of the kind was attempted by T. Marschak. He reduced the dimensions of the problem by studying the comparative efficiency of the centralized and decentralized frameworks. Marschak's results were not conclusive. He felt that the lessons which the current designer of new economic systems could draw from the Yugoslav experience were "staggeringly obscure" (1968, p. 586). Later research was undertaken in the Institute of Economic Studies (IES) (Horvat, 1969b). Efficiency was measured in terms of the rate of growth cf output attributable to technical progress, defined as the residual after the contributions of labor and capital have been accounted for. The results are summarized in Table 4.

The periodization in the table is not ideal and is determined by the availability of data. Yet the results of the analysis are extremely suggestive. In the foregoing section it was stated that the investment program of the First Five-Year Plan was completed by 1955. Statistical testing in the IES study showed that the Yugoslav economy operated on the basis of two completely different production functions, one applying to the period 1947–1955 and the other afterwards. The former had a negative residual, the latter a positive and a very large one. The table seems to suggest that central planning expanded output and employment fast, and capital formation even faster, as compared with the private capitalist pre-war economy. But it also reduced overall efficiency. Self-government

TABLE 4.—THE USE OF LABOR AND CAPITAL AND TECHNICAL PROGRESS IN YUGOSLAVIA

		GNP	Rates of growth per annum in %		Rates of growth of GNP due to increased efficiency
			Employ-ment	Fixed Assets	
Capitalism:	1911–1932	3.28	1.87	3.52	0.71
	1932–1940	4.67	0.72	2.59	3.16
Etatism:	1940–1954	5.91	4.76	9.99	−1.04
Self-government:	1956–1967	10.31	4.44	7.84	4.44

Note: The war years 1914–1918 and 1941–1945 are excluded. The data refer to manufacturing, mining, power generation, construction and crafts.

accelerated the growth of output and technical progress beyond anything known before while preserving fast employment expansion.

As might have been noticed already, the growth was fast but not at all smooth. At first the possibility of regular cyclical development was rejected by some economists. Yet in another IES study business cycles with periods of about four years were established (Horvat, 1970). These cycles, that manifest themselves as fluctuations in the rates of growth (see Figure 1), have interesting features not found elsewhere. Thus inventories are accumulated in the downswings and decumulated in the upswings; the accelerator is not operative; prices tend to vary inversely with the cycle etc. The upper turning points seem to be generated by divergent changes in import and export elasticities that end in an explosion of the balance of payments deficit. The lower turning points are somewhat more difficult to explain. Bajt believes that consumer demand is to a certain extent autonomous and helps to generate an acceleration of output growth (1969b).

If the beginnings of the cycles are measured from inflection points in the downswings of the rates of growth (these points correspond to peaks of deviations from an exponential trend of absolute magnitude), they appear to coincide with major economic reforms. Thus, the five cycles that have occurred so far describe in an interesting manner the history of post-war economic policy (Roman numerals indicate quarters):

1. Cycle: New Economic System (1), III/1949–III/1955.
2. Cycle: The transition to the Second Five-Year Plan, III/1955–II/1958.
3. Cycle: New System of Income Distribution, II/1958–IV/1960.
4. Cycle: New Economic System (2), IV/1960–L/1965.
5. Economic Reform, I/1965–?

Cyclical institutional development seems also to be a novel feature of business cycles.

Development Policy and Methods of Planning

Development Policy and Functions of Social Plans: The philosophy of development, generally accepted by Yugoslav economists and the Government until about 1956, is well described by Čobeljić, then the deputy director of the Federal Planning Bureau (1959a). Čobeljić maintains that rapid industrialization is the chief method of generating development. Industrialization creates additional urban employment, which alleviates latent un-

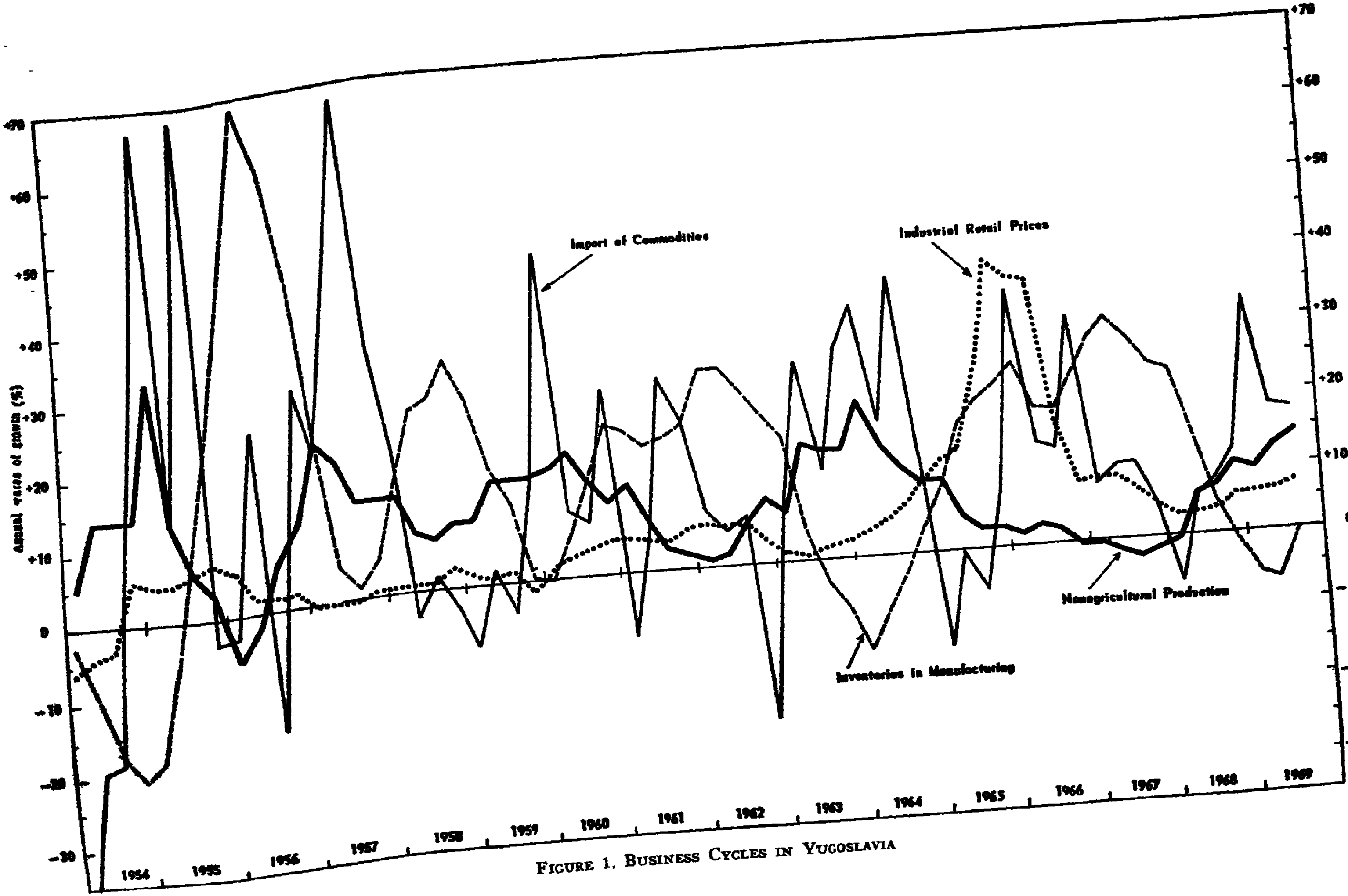


FIGURE 1. BUSINESS CYCLES IN YUGOSLAVIA

employment in agriculture. The growth of the urban labor force generates additional demand for agricultural products and so stimulates the development of agriculture. Physical control of foreign trade, in order to prevent the import of non-essential goods and to secure imports of capital goods, and the more rapid growth of consumer goods industries (so-called Department II) helps to accelerate industrial growth which in turn generates development impulses throughout the economy. Imports are paid for by exports of raw materials and agricultural products. The necessary saving is secured by a proper price policy. Prices in predominantly private agriculture are kept low and in consumer goods industries are inflated by means of high turnover taxes.

The policy described was not only advocated but was also consistently implemented. In the period 1950–1956 investment in industry (manufacturing, mining and power generation) absorbed 51 percent of all investment. The share of industry in national income rose from 21 percent in 1939 to 40 percent in 1956. Four fifths of industrial investment were channeled into heavy industry and power generation. The share of saving in national income increased four times as compared with the pre-war level (Čobeljić, 1959a, pp. 178, 366).

The planning system in 1947–1952 was consistent with such a development policy. The main characteristics of this system, as described by Jelić, another deputy director of the FPB, were as follows: a/ strict centralization of decisions about priorities, timing and structural changes; b/ physical allocation of resources as a basic method of planning; c/ financial elements play a secondary role and serve to achieve balances in value terms; d/ targets represent directives; e/ production is planned by commodities and capital formation is planned in terms of individual investment projects; f/ prices are administratively fixed; g/ the elements of which a plan is composed are also instruments of its implementation (Jelić, 1962, pp. 102–105).

After the engine of growth had been set into motion in the way described and the economy organized along socialist lines, there was a possibility of and a need for a different approach. Čobeljić now expected a more balanced growth. Jelić referred to Rostow's take-off theory and to Bićanić's threshold of growth theory (Bićanić, 1962b) and insisted on decentralized initiative as a further vehicle of growth. Self-government implied that the function of planning be separated from the function of operational management. Jelić pointed out that social plans should determine at least three global proportions—the basic division of national income, the structure of investment and the relations with foreign countries—if they were to be efficient devices for implementation of social preferences (Jelić, 1962, p. 144).

The same three global proportions were accepted as basic by D. Bjelogrlić, director of the Planning Bureau of Serbia. He added, however, a fourth one: the relative growth of the less developed states and regions (1965, p. 118). Bjelogrlić presented his paper to a conference on social planning held in Belgrade in 1965, where Čobeljić and K. Mihajlović spoke in favor of introducing more directives into planning, while M. Samardžija and M. Korač maintained that even the planning of the share of accumulation and the structure of investment meant a violation of self-government. This discussion, which covered a wide spectrum of opinions from semi-central planning to an almost complete *laisser faire* approach, has been characteristic of the Yugoslav economic profession since the enactment of the new Constitution in 1963. The trend has been towards the *laisser faire* extreme. In 1960 the Federal government controlled 48 per-

cent of business investment directly through its General Investment Fund, and in addition to that 14 percent indirectly through tied loans (Jelić, 1962, p. 158). In 1969, the Party Congress recommended that so-called state capital be eliminated and in the future the federal government is not supposed to retain any direct control over investment resources. A satisfactory solution has not as yet been found, and work on the new Law on Social Planning, which began in 1963, has not yet been completed (Savezna Skupština, 1966a).

The advocates of the new approach to planning, Sirotković (1966), the former director of the Planning Bureau of Croatia, and R. Štajner (Savezna, 1966a), the present director general of the Federal Planning Bureau, M. Mesarić (1967) and others argue that the professional function of planning should be supplemented by an emphasized social function, that annual plans should be abandoned and replaced by parliamentary resolutions (which has been practiced since 1966), and that medium-term plans should be continually revised and extended every two to three years. Bićanić describes the desirable system of planning as polycentric planning. This presumes the existence of one planning mechanism consisting of many autonomous plans interlinked in a specific, competitive way (Bičanić, 1963b, 1967). These ideas have been more or less accepted, but in parliamentary debates criticisms have been voiced that it was not at all clear how the plans were to be implemented (Savezna Skupština, 1966a, p. 91). In practice the implementation of plans has left much to be desired and the law on Social Planning is still to be produced.

The functions of social planning in the present Yugoslav setting have been described by the IES (Jugoslavenski Institut, 1968, p. 20), and similarly by Mesarič (1969), as follows: (1) A plan is, first of all, a forecasting device. (2) As such it provides economic subjects with necessary information for their autonomous decision-making. This, together with institutionalized consultations, makes the plan an instrument of coordination of economic decisions. (3) After relevant social preferences have been determined by an essentially political process, the application of modern tools of economic policy makes the plan an instrument for programming economic development. (4) Once the social Plan has been adopted by the Parliament, it becomes a directive for the Government. Point (4) is the only administrative or compulsory aspect of social planning.

Institutional Framework: The precondition for efficient social planning is an adequate analysis of the functioning of an institutional framework. A general idea of how the system works or is supposed to work may be obtained from a description by Horvat (1969c).

The Yugoslav economic system consists of autonomous, self-governing, work organizations[5] and individual producers in market and non-market sectors and of government machinery. The task of the latter is to use *non-administrative* means in coordinating the activities of market and non-market agents and to organize public administration in certain fields of common interest (judiciary, defense, foreign affairs, etc.).

The functioning of this economic system is based on the assumptions that the self-governing collectives are materially interested in maximizing their incomes and that the Government and Parliament are able to create an economic environment in which autonomous decision-makers be-

[5] "Working Organization" is a constitutional term meant to underline a fundamental equality in rights and status of every group of citizens organized with an intention to earn a living regardless of the activity they perform. An enterprise, a theatre and a government office—all of them are work organizations.

have in accordance with general social interests. Both assumptions seem to have been proved correct by the modern theory of economic policy and by experience in well organized market economies. Between the "Center" (Parliament) and the "Periphery" (Work Organizations) four types of gravitational forces are active in keeping the system in equilibrium and the economic agents on the predictable trajectories of social interest.

These forces are information—consultation ties, market ties, economic policy ties (instruments of economic policy and legislation) and administrative ties. The last mentioned are exceptional as far as economic agents are concerned and apply to various organs of the Center such as ministries, the National Bank, certain bureaus, and the like.

I should add that there is also a fifth type of ties—political ties—which closes the whole structure connecting the work organizations with the Parliament and with flows of commands (arrows) oriented from the Periphery towards the Center. In order to keep this section short, I shall not analyze these ties (this is why they are omitted from Figure 2). It is, however, important to realize that the Parliament is organized in a rather unorthodox fashion. Apart from the traditional Political House, whose members are more or less professional politicians, elected by all citizens, there are three additional houses, dealing with three different social–economic groups of problems (economic, health and welfare, education and culture). The members of these three "Houses of Work Unions" are not professionals; they keep their usual jobs and are elected by the "producers" in these three specific fields.

Let us now have a look at the market half of our economic cosmos. The activities of enterprises and individual producers are coordinated by the market in the first place. The market is, however, a very rough and unreliable mechanism requiring constant adjustments.

These adjustments are achieved through general regulative measures and the instruments of economic policy of the Government. The financial flows, intended to achieve a desirable allocation of resources, are regulated by the National Bank within the framework of the Social Plan. There are two additional types of specific financial interventions: in the field of foreign trade (credits and exchange risks insurance) and in investment (insuring proper structure and regional allocation of capital formation). These three purposes are served by three federal funds: for export credits, for underdeveloped regions and for investment.

Market equilibrium is being worked upon by three institutions. Two of them—the Directorates for food and for industrial products reserves—intervene whenever supply and demand do not match. The former Directorate also administers agricultural support prices. The third institution, the Price Control Bureau, is now a somewhat alien element in the system. I expect that in the near future this governmental bureau will evolve into a Price and Wage Arbitrator, an institution in which all relevant interests would be represented and all decisions made jointly. At the moment more than 40 percent of industrial prices are controlled.

Statistical and Planning Bureaus have only informative–consultative functions in this system.

A rather peculiar arrangement of the Yugoslav system is to be found in what I call a *quasi–market*. The activities of schools, hospitals, museums, and other non–market work organizations cannot be coordinated by the market directly as is done in the case of enterprises. In a socialist society sick persons should be healed, talented youths educated, regardless of whether and how much they can afford to

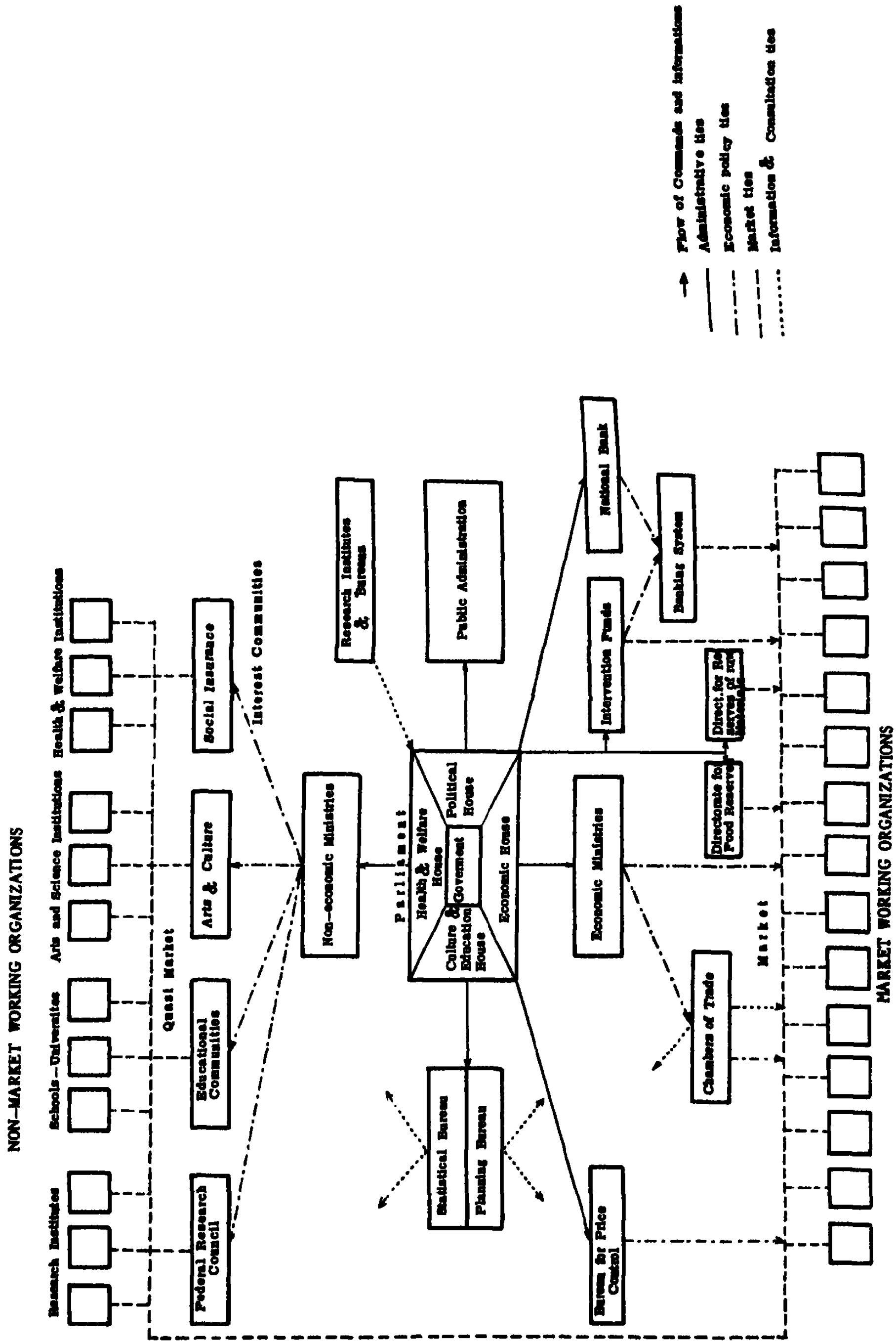


FIGURE 2. INSTITUTIONAL MODEL OF THE YUGOSLAV ECONOMIC SYSTEM

pay. On the other hand, the traditional budgetary financing of non-market activities has led to bureaucratic practices incompatible with a self-government system. The solution of this dilemma was sought in an interpolation of a special self-government mechanism between the government and the non-market working organizations. This mechanism is called interest unions. The unions obtain their financial resources on the basis of parliamentary decisions and then buy the services of non-market producers on behalf of the society. The non-market producers compete for available resources by offering their services on differential conditions. In this way, there emerges a special type of market–a quasi-market—which makes it possible for the relations between the non-market sector and the society to be economically conditioned, for the collectives in the non-market sector to preserve their self-government autonomy, and at the same time for relations within the fields of education, culture and social welfare to be based on the principle of "distribution according to needs," which is one of the preconditions of a socialist society. It is clear that the enterprises can also intervene in the non-market sector—either by buying services directly or by creating special foundations—and that is why in Figure 2 market and non-market sectors are also directly connected by market ties.

Apart from economic relations between federal bodies and economic agents there are relations between federal and state and local authorities, between the latter two and the work organizations and among all of them. I must, however, refrain from describing all these relations, although they are extremely important for the functioning of the system as a whole.

Other Issues: One of the recurring issues of the planning controversies is the problem of optimum investment and saving. Impressed by the unpleasant contraction of personal consumption during the Five-Year Plan, Čobeljić maintained that a certain minimum rate of growth of consumption represents the upper limit for the share of accumulation and for the growth of output (1959a, p. 188). Similarly, Stojanović argued that reduction of consumption growth below a certain limit reduces the growth of labor productivity and that this functional relation determines the optimum rate of investment (1960). Bajt also agreed that the optimum rate of investment is determined by the rate of consumption that maximizes the productivity of labor (1958), but this is not necessarily the socially desirable rate of investment. The purpose of production is to increase economic welfare, and a maximization of welfare through time can be ascertained only by a discounting procedure (Bajt, 1963). Horvat argued that pure time discounting is inconsistent because an individual will regret his present impatience at some later date, and utility discounting is impractical since it cannot be ascertained empirically. The other often suggested theoretical solution—social determination of the terminal stock of capital—is irrelevant, since no sensible planner ever insisted on carrying out a long-term plan. One constructs, say, twenty-year plans in order to take into account all relevant consequences of the decisions that are taken *now;* with every new element of information in time the plan is revised and the planning horizon pushed forward. The alternative approach suggested can be described as follows. Since every economy has a definite and very strictly limited capacity to absorb investment (in Yugoslavia the limit is around 35 percent of GNP, SNA definition), maximum growth is achieved when the marginal efficiency of investment is reduced to zero. If a lag of several months for achieving a certain level of consumption is an acceptable price

for maximizing consumption within one's lifetime, then the maximum rate of productive investment is identical to the optimum rate of saving. Thus maximization of the rate of growth appears to be a proper target for socialist planning. The trouble with the First Five–Year plan was not a low level of consumption but an inappropriately high level of investment. Pushed into the region of negative marginal efficiency, investment depressed output. A large part of such a stagnating output (up to 20 percent of national income) was used for defense. On both counts potential consumption was seriously reduced (Horvat, 1958, 1965).

Sophistication in economic analysis and planning methods has advanced considerably since the telephone age described by Stanovnik. Yet, both are still far from being impressive. Interindustry analysis has been adapted for planning purposes (Sekulić, 1968; Horvat, 1969d). Interindustry analysis was used in calculating the new exchange rate and the price levels in the last reform. Simple econometric models are now regularly used in the early stages of the preparation of a plan (Nikolić, 1964; Horvat, 1968b). An integrated system of social accounts, specially adapted for planning needs, has been produced recently (Horvat, 1968b; 1969e). For the rest, planners rely on abundant statistics, old–fashioned balancing and hunches. A satisfactory methodology of planning is yet to be written.

III. *Labor–Managed Enterprise*

Self–Management

Self–management is undoubtedly the most characteristic of Yugoslav institutions. Further developed into social self–government, it is the pivotal institution of the Yugoslav socioeconomic system. Moreover, Yugoslav social scientists are quite unanimous in believing that without self–government socialism is impossible (Fiamengo, 1965). Thus the fate of socialism depends on the feasibility and efficiency of self–government. In this section we will be concerned only with self–government as applied to business firms, which is usually denoted as self-management.

Self–management is not a Yugoslav invention. The development of this institution can be followed from the beginning of the last century (Horvat, 1969a, ch. 5). Every social revolution from the Paris Commune onwards attempted to implement the idea of self–management. In the very beginning of the revolution in Yugoslavia, in 1941, workers were assuming control over factories in various places (Tanić, 1963, p. 30). With the establishment of central planning, the idea of self–management suffered a setback. However, already in 1949 it was revived; by the end of that year workers' councils were created as advisory bodies in 215 major enterprises and in June 1950 the law passed that inaugurated the era of self–management.

For more than a decade the basic organizational principles of self–management remained unchanged. All workers and employees of a firm constitute the work collective (radni kolektiv). The collective elects a workers' council (radnički savet) by secret ballot. The council has 15 to 120 members elected originally for one year and recently for a two–year period. The council is a policy making body and meets at intervals of one to two months. The council elects a managing board (upravni odbor) as its executive organ; the board has 3 to 11 members, three quarters of whom must be production workers. The director is the chief executive and is an ex officio member of the managing board.

As soon as it was established, self–management met with criticism and skepticism. Both came mostly from abroad. It was said that self–management would erode discipline and that workers would distribute all profits in wages, thus reducing the

growth potential of the economy. In 1955 Ward suggested that workers had no real choice in the election of the council and that actions reportedly taken by the councils might represent rubber stamping (Ward, 1957; Horvat and Rašković, 1959). In evaluation of these criticisms one may point out that, regarding labor discipline, an International Labor Organization mission found in 1960 that "... while the self-government machinery for labor relations has curtailed the former powers of the supervisory staffs, it would not appear to have impaired their authority. . . . It has undoubtedly strengthened the position of the collective vis-à-vis the management, but it does not appear to have undermined labor discipline" (International, 1962, p. 203). As to the growth potentials, the rate of accumulation remained high with a chronic tendency towards overinvestment and with a high rate of growth. Elections are supervised by courts, and all candidates approved by the majority of the workers are included in the voting list. The safeguards against the creation of a managerial class are the workers' majority in the managing board and the provision that members of self-managing bodies may be elected only twice in succession.

The real difficulties were encountered elsewhere. The original organizational scheme proved to be too rigid, and had to be revised extensively in all its three components. It soon became evident that the director's position was not quite compatible with the new arrangement, and directors came to be "one of the most attacked and criticized professions in the country" (Novak, 1967, p. 137). In the étatist period the director was a civil servant and a government official within the enterprise. He was in charge of all affairs in the enterprise and responsible exclusively to the superior government agency. In the self-management system the director became an executive officer of the self-management bodies, while at the same time continuing to represent the so-called public interest in the enterprise. This hybrid position has been a constant source of conflicts. At first the director was appointed by government bodies. In 1952 the power of appointment of directors was vested in the commune. In 1953 public competition for the director's office was introduced and in the selection committee the representatives of the commune retained a two-thirds majority. In 1958 workers' councils achieved parity with communal authorities on the joint committees authorized to appoint and dismiss directors of the enterprises. The present state of affairs is that the director is appointed by the workers' council from candidates approved by the selection committee on the basis of public competition. He is subject to re-election every four years, but may also be dismissed by the workers' council. Since the appointment of the director does not depend exclusively on the will of the collective–as is the case with all other executives—he has been considered a representative of "alien" interests in the firm. There have been constant attempts to reduce his power, which have made his position ambivalent and reduced his operational efficiency. On the other hand, as G. Leman remarks, the director is expected to play the triple role of a local politician, a manager and an executive (1969, p. 28). In the context of what has just been said, the managing board was supposed to exercise control over the work of the director and the administration. Involved in problems of technical management and composed of nonprofessionals, the managing board often proved to be either a nuisance or ineffective. For professional management the director had to rely on the college of executive heads (kolegij), which was his advisory body and subordinated to him. Thus two funda-

mentally different organizational setups were mechanically fused into one system. The director's office provided a link between them, i.e. between the self-management organs and the traditional administrative hierarchy.

Finally, in any somewhat larger firm one single workers' council was not sufficient if there was to be real self-management. In 1956 workers' councils on the plant and lower levels were created apart from the central workers' council. Even this was not sufficient, because hierarchical relations between workers' councils at various levels were not compatible with the spirit of self-management. "The self-management relation in its pure form is polyarchic and not democratic"—explains D. Gorupić—"the democratic relationship represents a domination of the majority over the minority. . . . The polyarchic character of the self-management relationship is revealed in equal rights of members of a certain community" (Gorupić, 1969, p. 16).

In 1959 an interesting new development began with the creation of so called economic units (ekonomske jedinice). The enterprises were subdivided into smaller units with a score or several scores of workers. Since a year earlier the enterprises had become more or less autonomous in the internal division of income, it was thought that a strong incentive could be built into the system if economic units recorded their costs, took care of the quality of output, use and maintenance of machinery, and themselves distributed their incomes according to certain efficiency criteria. In an interesting study Lêman, a German student of Yugoslav self-management, argues that economic units resulted from endeavoring to eliminate dividing lines between three fields of activities: policy making, managing and executive work (1967, pp. 38–39). Soon, economic units began to practice collective decision making on all sorts of matters. It became advisable to enlarge economic units so as to comprise individual stages of the technological process or separate services. Economic units were transformed into work units (radne jedinice). The hierarchical self-management relations within the enterprise called for a revision. Important self-management rights (distribution of income, employment and dismissals, assignment to jobs) were transferred to work units. Direct decision making at meetings of all members of the work unit became the fundamental form of management. In this way the work unit provided a link between the primary group and social organization. It was both a well defined techno-economic unit, meeting the requirements of efficient formal coordination, and the basic cell of workers' self-government (Županov, 1962).

Work units, several workers' councils and managing boards, many commissions and committees—all this made the formal organization of a labor-managed enterprise rather complicated and inefficient. In order to make such a formal system work, it had to be simplified in practice and this was done in various informal ways. That in turn meant further limitations on competent professional management and a further reduction of efficiency. Workers' management is passing through an efficiency crisis caused by the need for a radical transformation of inherited organizational structures. After all, workers' management meant a fundamentally new principle in running enterprises and it would have been surprising if that did not require painful adaptations and deep changes in social relations. I must add, however, that the conclusions in this paragraph, though based on widely held beliefs, cannot be substantiated in a more rigorous way because no adequate empirical research has been undertaken so far.

Although the crisis has not yet been overcome, matters have begun to be gradually sorted out. A constitutional amendment, passed in 1969, made it possible for enterprises to drop managing boards and to experiment with various organizational schemes. Trade unions, authorities and workers have come to realize that certain developments were based on erroneous beliefs concerning various management functions in a labor-managed enterprise. Perhaps the clearest analysis of the mistakes made came from a sociologist, J. Županov (1967a). Županov distinguishes self-management (samoupravljanje), management (upravljanje) and executive work (rukovodjenje). The last mentioned is a partial activity intended to carry out a decision made within a policy framework. The integration of all decisions into a consistent framework is the task of management. But management means only technical coordination, while coordination of various interests, making basic policy decisions, is a task of self-management. Self-management means social integration, the formulation of common goals, which is a precondition for efficient operational work of the management. The confusion between management and self-management generated tendencies to transfer more and more of formal coordination to bodies whose task was social integration. As a consequence, satisfactory social integration was not achieved, while non-professional management meant lower efficiency (Bilandžic, 1969). S. Bolčić has reminded me that this inherently complex problem was complicated even further by a rather naive ideology contained in legislation and political propaganda and advocating direct participation in administrative work as an indispensable of safe-guarding the interests of the workers.

How are the problems encountered to be solved?

Gorupič (1967) and the IES (Institut, 1968b) saw the solution in a fusion of professional competence and self-management. The enterprise may be considered an association of work units. The professional managers of the work units should no longer be appointed, as in the traditional set-up, but be elected by their associates. In this way they would represent the interests of their primary groups, while at the same time being also professionally competent. Managers so elected would make up a managing board which would be both an executive organ of the workers' council and a professional management body. Decisions would be made collectively. Since most of the decisions affecting the daily lives of workers would be made and implemented within economic units and by themselves, executive work would become more and more purely organizational and lose its order-giving character (Novak, 1967, p. 118). Businessmen proved susceptible to this approach (Miletić, 1969). As one might have expected in a country like Yugoslavia, as soon as these ideas had been clearly formulated the practical experimentation began, and the Constitution was promptly amended.

Before closing this section let me note another interesting phenomenon: the development of the so-called autonomous law. Enterprises appear as law-creating bodies. Their self-management organs pass charters and rules governing the organization of work, the composition and responsibility of self-management and other organs, the distribution of income, and the conduct of business. The autonomous law-creating power emanates directly from the Constitution, the rules and regulations are legally binding on all persons to whom they are addressed within an enterprise and disputes are settled by the enterprise organs, except in some specific cases. In this way "a continual narrowing of the area of state law and corresponding broadening of the area of so called autonomous law

characterizes the entire process of regulation of social relations in Yugoslavia" (Kovačević, 1969, p. 1).

Enterprise

The introduction of self-management in 1950 implied the dissolution of the centrally planned, administratively run economy. The enterprise was to become independent and autonomous. Individual enterprises needed some guidance and coordination. Therefore so called Higher Business Associations (viša privredna udruženja) were set up in order to replace former state directorates and to preserve continuity in the organization of the economy. The governing councils of the new bodies were composed of representatives of workers' councils of the constituent enterprises. But Higher Business Associations tended to operate along the same administrative lines as former directorates and were therefore dissolved in 1952. A period of *laissez faire* ideology followed. Isolated enterprises were expected to engage in free competition on the market. Attempts to form larger business units and multiplant firms were frowned upon as contrary to genuine self-management and as signs of going back to a disguised state control. In spite of that the system worked well because a special sort of administrative coordination was still effective. The chief coordinator was the Bank implementing the targets of the Plan. The Bank operated a specially designed bookkeeping for every enterprise, distributed the incoming money to various accounts (for wages, taxes, and various enterprise funds) and determined the amount of the necessary working capital which was to be provided on a credit basis, etc. (Vučković, 1952, pp. 11–2p). Although the control was monetary, the value proportions were derived from physical targets.

After 1952 the process of decentralization was not arrested at the level of the enterprise, but went below it. It has already been mentioned that in 1956 the formation of plant workers' councils began and in 1959 the first economic units appeared. The internal cohesion of the enterprise was reduced and it looked as if it was broken up into its component parts. At the same time various monetary and non-monetary administrative controls were gradually being removed. In 1954 the enterprise assumed control of its fixed capital. Fixed assets could be bought and sold without asking for permission. Investment auctions were tried out. In 1958 the enterprise gained control over the internal distribution of income and two years later the trade union control of wages was removed. The stage was set up for a genuine market economy.

As soon as all preconditions for classical free competition of numerous small enterprises were met, it became clear that such an economy would not work very efficiently in the second half of the twentieth century. Since the state refrained more and more from coordinating economic activities, some other agency or agencies had to replace it in that function. That is why the process of integration was initiated. Working collectives themselves had to resume economic coordination in a state that was withering away. The circle of organizational development seemed closed. The process was started by a fully integrated state managed economy, passed through a period of radical decentralization and is now moving towards another stage of full integration in the form of a labor-managed economy.

The forms of integration are various. The simplest one is an agreement for business cooperation intended, for instance, to achieve specialization of the production programs of two or more enterprises. Next comes contractual techno-economic cooperation resulting in joint production, sales or procurement of raw materials. If

business relations are numerous and complicated so that it is not possible to regulate everything in advance in a contract, the enterprises form a separate body called a Business Association (poslovno udruženje). By 1962 already one half of manufacturing enterprises were members of Business Associations that first appeared in 1958. In 1967 there were 290 Business Associations consisting on the average of ten enterprises (Dautović, 1968). The next more integrated form is a firm called Affiliated Enterprises (združeno poduzeće). Such a firm is run according to commonly accepted business principles, while constituent enterprises retain operational independence. The latter disappear in a merger. In a seven-year period, starting with 1959 when the process began, the total number of firms was reduced by one half by mergers. It is characteristic, however, that nine-tenths of these mergers were effected within the boundaries of the same or neighboring communes, and only 1.2 percent were interestate mergers. In the same period the number of banks was reduced from 378 to 108. Special status was given to so-called *Unions of Enterprises* (zajednice privrednih organizacija) created for railways, electric power generation and postal and communication services. Membership in these Unions is obligatory. Finally, there are Economic Chambers, organized territorially and associated in the Federal Economic Chamber. The Chambers have a dual role: they help their members in various ways and they also perform a public function, mediating between the state and the business interests. Membership is obligatory.

Mergers and various forms of business cooperation may mean monopoly. That is why a sort of anti-monopoly legislation appeared as well. It is explicitly forbidden to limit free competition in production or sales to any enterprise outside the business group concerned, and government inspectors are expected to take care that there is no sharing of the market or connivance about prices. No serious research about possible monopoly practices has been undertaken as yet, and so there is no possibility of presenting an evaluation here. But it must be borne in mind that the Yugoslav economy will behave differently from other market economies. Workers' management implies a spontaneous public supervision of business conduct and so classical forms of collusion, characteristic of private monopolies, are hardly to be expected, J. Dirlam (U.S. Congress, 1968, p. 3854) finds that the degree of output concentration is higher in Yugoslavia than in the United States; J Drutter (1964) establishes the non-existence of correlation between profits and output concentration and similarly H. Wachtel (1969) finds no correlation between wages and output concentration. In spite of a considerable number of mergers in the period 1959–1963, the degree of concentration actually decreased (Tanić, 1963).

A new enterprise may be founded by an already existing enterprise, by a government agency or by a group of citizens. The founder appoints the director and finances the construction. Once completed, the enterprise is handed over to the work collective which elects management bodies. As long as all obligations are met, neither the founder nor the government have any say about the operations of the enterprise. Enterprises are also to merge or to break in parts. If a work unit wants to leave the mother enterprise, and the central workers' council opposes that, a mixed arbitration board composed of representatives of the enterprise and of the communal authorities is set up. In all these cases it is, of course, implied that mutual financial obligations will be settled.

Since the capital of an enterprise is socially owned, the fundamental obligation of the enterprise is to keep capital in-

tact. If it fails to do so for more than a year, if it runs losses or fails to pay out wages higher than the legal minimum for more than a year, the enterprise is declared bankrupt or the founder undertakes to improve its business record. In the latter case self-management is suspended and replaced by Compulsory Management (prinudna uprava), a form of receivership administered by officials chosen by the commune (Miljević, 1965). Bankruptcies are rather rare because the commune is obliged to find new employment for workers and so prefers to help the enterprise as long as possible.

If integration processes are to proceed efficiently, the organizational forms must be extremely flexible. Thus since 1967 it became legally possible for two or more enterprises to invest in another enterprise and then share in profits. Similar arrangements were adopted in joint ventures with foreign capital (Friedmann and Mates, 1968; Sukijasović and Vujačić, 1968). In an open economy, like the Yugoslav one, foreign capital is welcome provided it does not limit workers' self-management. Therefore direct investment is impossible, but joint ventures are encouraged. The basic motivation for a Yugoslav firm to enter into close business cooperation with a foreign partner is to be found in the desire to secure access to the knowhow and the sales organization of the foreign firm. In this way the Yugoslav firm tries to achieve international standards in technological efficiency and to expand its market.

Theoretical analysis of the behavior of the Yugoslav firm has only begun. Oddly —or understandably—enough, the pioneering work was done by a foreigner, B. Ward of the University of California at Berkeley. In his 1958 paper on the "Illyrian" firm (1958), Ward argues that rational behavior will require maximization of income per worker. In the Marshallian short-run, one product, one factor case this leads to some queer consequences: an increase in wages leaves output and employment unchanged, an increase in fixed costs increases output and employment, and an increase of product price reduces output and employment. In a similar analysis eight years later, Domar showed that by generalizing the production function to include several products and several factors and by introducing the demand curve for labor the results are changed and begin to resemble the traditional conclusions about the behavior of the firm (Domar, 1966). Proceeding along similar lines D. Dubravčić comes to the conclusion that in a labor-managed firm there will be a strong tendency to use capital intensive technology (1967). The empirical evidence does not give unequivocal support to this conclusion. While on the one hand there is a chronic hunger for capital and enterprises use every opportunity to invest, Yugoslav enterprises are also full of redundant workers. Instead of postulating what should be rational, the present author observes the actual practice of Yugoslav enterprises which fix wages in advance for the current year, and at least once a year make corrections (positive or negative) depending on the income earned. If this behavioral rule is used in the analysis, the results are again the same as in the traditional theory of the firm (Dubravčić, 1968). The last in this controversy, Dubravčić, points out that comparative analysis is really not legitimate because it is assumed that a capitalist firm maximizes an absolute magnitude (profit) while a socialist firm is expected to maximize a relative magnitude (income per worker). Dubravčić suggests a symmetrical treatment on the basis of the entrepreneurial input, which is capital in the capitalist case and labor in the socialist case. If a capitalist firm maximizes the rate of profit (profit per unit of capital) it will behave in ex-

actly the same way as Ward's Illyrian firm with entrepreneurial inputs being interchanged. In both cases firms will economize on the entrepreneurial input and this will lead to capital intensive techniques in a socialist firm and to labor intensive techniques in a capitalist firm (Horvat, 1967a)—a nice and almost humorous result.

This brings us to the problem of entrepreneurship in a labor-managed firm. If an entrepreneur is a risk taking and innovating agent—as Knight and Schumpeter would say and most economists would agree—then the work collective qualifies for that role (Horvat, 1964, ch. 6). In fact the work collective is generally treated as an entrepreneur. However, doubts have been voiced as well. Županov argues that the practice of fixing wages in advance means that they are not a residual in the income distribution—as is profit in a capitalist firm—and that this sets up a barrier to entrepreneurial behaviour. He quotes results of empirical research according to which in work units only managers and professionals are prepared to bear risks, while other categories of workers and employees mostly are not. S. Bolčić has drawn my attention to the fact that workers behaved rationally if they were prepared to bear risks only to the extent that they were able to control business operations. That is why managers were both prepared and expected by others to bear risks to a much larger extent. Such as explanation was spelled out explicitly by workers in a case quoted by Lêman (1969, p. 40). In another piece of research undertaken in Zagreb in 1968 it was found that all groups were more prepared to share in losses if output was diminished than if income was reduced while output remained the same or even expanded (Županov, 1967b). On the other hand, it is an empirical fact that wages vary pretty widely depending on the business results. Wachtel quotes data on the issues discussed at workers' councils meetings: two thirds of the agenda items are concerned with general management issues (labor productivity, sales, investment, cooperation with other enterprises, work of management) and only one third with direct worker issues (personal income, vocational training, fringe benefits) (1969, p. 58). Variable wages derived from profits amount to 8–14 percent of standard wages on the average (Wachtel, 1969, p. 100).

The Ownership Controversy

In Marxist sociology ownership relations are the basic determinants of social relations and thus of the socio-economic system. The class that owns—i.e. has an economic control over—the means of production, rules the society. For a long time, and in most instances even today, it has been maintained that private property generates capitalism and state property socialism. In fact the percentage of the national capital owned by the state has been taken as the most reliable measure of the degree of socialism achieved. It follows that a socialist economic policy must be oriented towards an overall economic control by the state and must be hostile towards private initiative.

As already noted, the above described view was generally accepted in Yugoslavia until 1950, and since then it has been thoroughly revised. It is now pointed out that there are at least three reasons why the dogma of the identities between private ownership and capitalism, and state ownership and socialism, is false: the artisans of medieval towns were private owners but not capitalists; in ancient Oriental kingdoms state ownership was frequent and yet that had nothing to do with socialism; in fascist countries the state extensively controlled social and economic life while these countries were obviously capitalist (Horvat, 1969a, Ch. IV). Yugoslav scien-

tists are now quite unanimous in believing that state ownership may be a useful device to initiate socialist reconstruction, but is otherwise as alien to socialism as is private ownership. The present position is well summed up by J. Djordjević (1966, pp. 81, 79): "... state ownership of means of production creates a monopoly of economic and political power and ... makes possible the unification of economic and political power under the control of a social group personifying the state." Thus "... the essence of classical (class) ownership is not changed... As the holder of the title to property, it (the state) disposes with the producers' labor and its results, on the basis of which surplus labor is appropriated by groups which have their own interests in keeping their commanding functions and thus retaining power and their social status and prestige."

If state ownership fails to promote socialism, what is a feasible alternative? The Yugoslav answer is: social ownership. But the answer to the next question—What precisely is social ownership?—is not so easy and simple. The legal experts agree that social ownership implies self-government, that it is a new social category, that, if it is a legal concept, it does not imply an unlimited right over things characteristic of the classical concept of property, and that it includes property elements of both public and private law (Toroman, 1965, p. 5). In practically everything else there is disagreement. A. Gams and a number of other writers maintain that social property also implies rights of property since property implies appropriation, enterprises are juridical persons and the basic ingredient of the juridical person is property (Gams, 1965, p. 61). Article 8 of the Constitution says that the disposal of means of production in social ownership and other rights over things will be determined by the law. S. Pejović talks about the right of use which is somewhat wider than *usus fructus*, because it makes possible the sale of capital goods, but is narrower than ownership because the right of disposal is not absolute (Pejović, 1966, p. 29). A diametrically opposite view is expressed by Djordjević, and most other writers who maintain that social property represents a negation of property rights (Djordjević, 1966, pp. 84, 90). Djordjević quotes Part II of the Basic Principles of the Constitution to support his view: "Since no one has the right of ownership of the social means of production, no one—neither the socio-political community[6] nor the work organization nor an individual working man—may appropriate on any property-legal ground the product of social labor, or manage and dispose of the social means of production and labor, nor can they arbitrarily determine the conditions of distribution."

Legal writers differ further according to whether they stress the public law or private law component of social property. Further disagreements relate to the subjects of law (state, society as a real community of people, several subjects, no subjects). Next come disagreements on whether social property is a legal, economic or sociological concept or is non-definable in these terms because it relates to quasi-property. And if it is a legal concept, it may be so in various ways. By applying the calculus of combinations we can easily determine the number of possible theories. It seems that available possibilities have been efficiently exploited since M. Toroman (1965) was able to describe thirteen different theories.

The legalistic controversy was somewhat less interesting than the one among economists and sociologists that followed.

[6] Territorial political unit such as a commune, a district, an autonomous province, a republic and the federation.

Bajt drew attention to the fact that the legal owner and economic owner may be two different persons. The former holds legal title, the latter derives the actual benefit from the use of a thing (Bajt, 1968). In this sense social ownership implies the non-existence of exploitation which in turn implies the distribution of income according to work performed. If a person or a group of persons are earning non-labor income, they are exploiting others, and in so far as this happens social property is transformed into private property. Thus self-management per se is not a sufficient condition for the existence of social property.

The institution of property already undergoes gradual disintegration under capitalism. Shareholders are legal owners but management exerts real economic control. That is why I prefer to replace the traditional concept of property by a more fundamental concept of economic control (Horvat, 1969a, Ch. 15). The latter always means "control over labor and its products" which is Marx's definition of capital as a social relation (Marx, 1953, p. 167). In this respect legal titles are irrelevant. If artisans or peasants possess no monopoly power—which in an orderly market system is likely to be the case—then they represent no alien elements in a socialist society. And there can be little double that they practice self-management. Horvat and Bajt came to the conclusion that individual initiative is not only compatible with but is an integral part of a socialist system. In fact the process of production can be organized individually or collectively and that is why Bajt talks about two forms of social ownership: individual and collective.

Agreement about the matters mentioned so far is quite universal by now. Differences in views appear when intermediate cases are considered. Yugoslav law makes it possible for artisans and inn keepers to employ 3–5 workers. V. Rašković (1967a, pp. 106–107) and many others consider this to be a form of exploitation, a remnant of the old society, something alien to the system but which has to be tolerated at the present level of development. In support of this view Rašković argues that the employer would not hire workers if this were not profitable for him. It may, however, be argued in reply that a worker, by choosing an individual employer instead of a firm, reveals that he finds such employment more profitable for himself. Such a line of reasoning leads clearly to an impasse. To resolve the question whether workers may be hired by individual employers, and if so how many of them a sociological argument has been advanced as a criterion. As long as an individual employer works himself in the same way as his employees and has not become an entrepreneur merely organizing the work of others, employees may be considered as (often younger) associates in the work process, direct personal relations of a primary group are preserved and the alienation phenomena of wage labor relations are not present.

Discussion of the scope and role of individual work was invited by political bodies and very soon decisions were made following more or less the ideas expounded above. Individually organized production became a constituent part of a socialist economy.[7]

IV. *Market and Prices*

Price Policy

Price Policy represents an incessant series of attempts to control the famous law of value (supply and demand relations). Its history is instructive since it

[7] The private sector—which Yugoslav economists prefer to call "individual sector" in order to avoid various connotations of the attribute "private"—accounts for 29 percent of GNP and this percentage has not changed in the last fifteen years.

provides an insight into the working of various institutional arrangements.

Administratively Set Prices: Immediately after the war, with the economy almost totally destroyed, there was an extreme scarcity of all goods. The prime purpose of economic policy was to prevent profiteering and to generate output by any means available. This was the period of "profitability at all costs" (Radulović, 1968, p. 143).

Prices were determined on the free market only for a few luxury products. Mostly prices were set on the basis of actual costs incurred and could vary from one producer to the other. The Price Offices would examine each case and make the relevant decisions (normiranje cijena). This was not a very efficient procedure. Since actual cost was taken as given, there was no incentive to economize on inputs. Wages were fixed, and products could always be sold. In order to minimize risk, producers tended to inflate costs in their price proposals and in order to keep prices down the Price Offices tended to apply linear reductions to proposed prices. The authorities and the businessmen began to play at hide-and-seek, which is so characteristic for an administratively controlled economy.

The launching of the First Five-Year Plan in 1947 required a system of uniform prices (jedinstvene cijene). Uniform prices were determined by the planning authorities and were expected to be rigidly stable. The aim was to provide a link between the physical and the value part of the plan, to have a control over the implementation of plans and to avoid the administrative costs of changing prices frequently. Prices were formed by adding an average rate of profit to average cost for a product. The less efficient producers had planned losses, the more efficient ones extra-profits; in both cases differences were settled with the budget. Through the establishment of the system of uniform prices, the law of value was considered to be subject to an efficient social control (Kidrič, 1948, p. 143).

It soon became evident that uniform prices did not equilibrate supply and demand. There was chronic excess demand. Private producers (peasants and artisans) held a large share of the market and their incomes could not be easily controlled. Most consumer goods were rationed and sold at the existing uniform prices, but available quantities of consumer goods were not sufficient to satisfy the needs of the entire population at the lower uniform prices. By the end of 1947 the first quantities of consumer goods were supplied to the free market at higher uniform prices (više jedinstvene cijene). These prices were derived from the existing uniform prices by applying multiplying factors varying from 2 (for potato and beans) to 6½ (for garments). The resulting trading profit was absorbed by the budget. In 1948 about 45 percent of consumer goods were supplied at higher uniform prices (Šefer, 1956, p. 376). In this way, it was hoped, excess money incomes would be absorbed.

In agriculture a system of compulsory deliveries (obavezni otkup) was applied. Peasants were obliged to sell most of their products to the state at prescribed low prices. For the money they obtained they could not buy all those industrial products they wanted. Thus they tried to reduce deliveries and substitute their own consumption for money incomes. The government reacted by creating a market for industrial goods at higher uniform prices. Peasants reciprocated by evading compulsory deliveries and supplying more goods to the free peasant market, the only section of the market where the prices were equilibrating supply and demand. These prices tended to rise fast and so the government decided to substitute a carrot for the stick: in 1948 the government introduced linked prices (vezane cijene). Agricultural

prices were linked with industrial prices in such a way as to establish the pre-war parity. Peasants sold their products to the government at lower prices and in return obtained coupons which enabled them to buy industrial products at prices that were about 16 percent lower than commercial prices (Dobrinčić *et al.*, 1951, p. 141).

Local markets were less rigidly controlled. After 1949 local enterprises could in principle sell their products at commercial (higher uniform) prices. Trading establishments that were supplied by two different producers—national and local—were now unable to sell commodities at one single price. And so sliding prices (klizave cene) were invented. The selling price slides in a span determined by the lowest and the highest supply price. These prices were approved by the local authorities. Thus two different markets were created: one for enterprises that traded at lower and higher uniform prices, and the other for retail trade and population where prices approached free market prices.

The system of linked prices did not work too well. The supply of industrial goods was inadequte and richer peasants began to speculate with coupons. In 1950 only some agricultural products could be sold at linked prices. More of the peasants' products went to the free peasant market whose counterpart in the state sector was the system of sliding prices. Higher uniform prices, being administratively set, were lagging behind the free market prices. Output of consumer goods was stagnating, even falling, while incomes were rising (Čobeljić, Mihajloić and Djurović, 1954, p. 49): (see table 4A).

The widening gap between supply and demand could be controlled by administrative or economic means. The government chose the latter. In the transitional year of 1951 there were eight different price categories coexisting simultaneously (Dobrinčić *et al.*, 1951, p. 143). Sliding prices were superseded by higher prices for consumer goods. Rationing was abolished. Consumer goods prices were left to be regulated by the market while producer goods prices were increased one to twelve times and then frozen for about half a year. In 1952 compulsory deliveries of agricultural products were abolished. By the second half of 1952 all prices were freely formed with the exception of a few goods (bread, sugar, electric power etc.) for which ceiling prices were established.

Development of the Market: The strategy of the 1951/1952 price reform can be summarized as follows: (*a*) a sufficiently large increase of prices to absorb all excess money incomes; (*b*) a sufficiently large increase of retail prices of manufactured consumer goods relative to agricultural prices to generate the capital accumulation necessary for fast growth, (*c*) a smaller increase in producer goods prices to stimulate investment and the expansion of compartment I (producer goods industries). The first two goals were achieved with remarkable success. As a result, industrial producer prices were kept stable over a period as long as a decade. The third strategy proved to be deficient and generated a lot of trouble.

While the general index of industrial producer prices was declining for almost three years, prices of certain raw materials (ferrous and nonferrous metallurgy, building materials, wood products were rising.

TABLE 4A

	1948	1949	1950	1951	1952
Consumer purchasing potential	100	128	125	245	327
Retail trade in real terms excluding peasant trade	100	100	94	70	77

That is why in 1954 ceiling prices were set by the government for a number of raw materials, and in the next year the list of controlled intermediate goods was further extended. In 1955 industrial producers prices rose by five percent, which led to the creation of the Federal Price Office in the same year. Since then a system of administrative control of prices has been gradually developed. The essential features of this control are as follows:

1. The government sets *fixed* prices for electrical power, cigarettes, transportation rates, sugar, oil, salt and some other commodities.
2. The government sets *ceiling* prices for metallurgical products, coal, petroleum and some other goods.

These two categories of prices are changed at infrequent intervals. But when they are changed, the change is rather drastic.

3. Control on the basis of *prior price registration* is the most frequent kind of control. It was introduced in 1958. Producers intending to raise prices are obliged to notify the Federal Price Bureau thirty days beforehand. If within this period the FPB does not veto the price increase, it can be effected. The principal criteria for placing a product under control are: (a) its importance for the standard of living or for production costs of other products; (b) scarcity on the market and (c) the monopoly position of the producer (Vuković, 1968).
4. *Control of trade margins* is implemented by republics for wholesalers and by local authorities for retailers.
5. *Price freeze.* This instrument was used only on two occasions, in 1952 and in 1965, during two price reforms.
6. Agricultural prices are placed under a special regime. *Guaranteed prices* are applied to staple food products. This means that the Federal Food Reserve Board is obliged to purchase all quantities of the products offered for sale and to pay guaranteed prices. For milk and industrial crops *minimum* prices apply. This means that if these products are bought at least the minimum prices have to be paid for them. An industrial crop is normally not grown unless the producer has a prior contract with the buyer. Prices used in such cases are *agreed–upon prices.*

Industrial prices have been most heavily controlled. In the last decade this control was exercised over the following percentages of the value of industrial output (Radulović, 1968, p. 282; Drutter, 1968, p. 113; Institut, 1969, p. 6):

TABLE 5

1958	31.2	1965	70
1962	67.0	1966	66
1962–65	60.0	1967	53
		1968	46

The time series of prices, given in Table 6 may give an idea of how efficient the price policy and price control were.

After 1961 the administrative control of prices was increased and so was the inflationary pressure. What in fact happened?

The most frequent form of price control—prior price registration—could not be adequately applied to new products. By making small changes in the design of a product an enterprise would transform it into a new product and so could evade price control. In 1964 almost twenty five thousand new products were launched. Low and rigidly controlled prices of raw materials made their production unprofitable and so depressed output; in agriculture prices were particularly depressed. That is why in 1964 prices were raised administratively in agriculture, the food processing industry, energy generation and nonferrous metallurgy. Next, differential taxation, a system of premiums and subsidies, and administrative interventions in foreign trade tended to preserve and even increase price disparities (Pertot, 1966). As

TABLE 6.—CHANGES IN PRICE LEVELS IN PERCENTAGES PER YEAR

	1952–1963	1964	1965	1966	1967	1968
Producer prices in manufacturing and mining	+0.9	+5	+15	+11	+2	0
Agricultural producer prices	+8.6	+24	+43	+16	−3	−4
Retail prices (including services)	+3.9	+9	+29	+23	+7	+4

Sources: *Jugoslavija 1945–1964*. *SGS–1969*.

a consequence individual enterprises conducted their business under highly unequal conditions. Producers whose prices or wages were lagging behind were trying to catch up with their neighbors. The Federal Price Bureau received 12,800 requests for price increases in 1961 and 69,000 requests in 1964 (Drutter, 1968, p. 107). But the most important reason for the break in price trends in 1961 lies elsewhere. Until 1961 personal incomes were quite efficiently controlled by fiscal and nonfiscal means (Trade Unions). That is why prices were quite stable (except in agriculture) and administrative controls relatively few (Institut, 1968b, pp. 37–41). In 1961 income controls were abolished, very soon a cost–push inflation occurred and, despite increasing administrative control, prices went up. A few years later the Institute of Economic Studies suggested that the Federal Price Bureau relax administrative price control and focus its attention on income control (Institut, 1969, p. 41). The suggestion was not followed and instead monetary policy was used as the chief anti-inflationary weapon.

By 1965 the economy was ripe for another radical price reform. In March prices were frozen and a tax reform carried out. Various subsidies were drastically reduced and the tax burden of enterprises alleviated. In the next few months a new price structure was prepared. In July the dinar was devalued; new prices were introduced and frozen. Relative prices of certain raw materials, intermediate goods (electric power, petroleum, ferrous and nonferrous ores and metals, chemicals, timber products and agricultural products) and transportation services were substantially increased. World prices (as registered in exports or imports) were taken as a basis for the new price structure. This was to make possible a rapid integration of the Yugoslav economy into the world economy. World prices were corrected upwards or downwards by taking into account capital accumulation needs of various industries and other specific purposes. A new customs tariff was to iron out these differences.

The price stabilization proceeded rather slowly, as can be seen from Table 6. The lifting of price controls went even slower. In 1968 prices looked stabilized, but almost one half of industrial prices were still under control. Disparities between controlled and uncontrolled prices began to emerge. The output of certain industries tended to become depressed. In 1969 prices began to rise again. The experience of 1964 seems to have been repeated. The reform of 1965 eliminated the worst price disparities, but subsequent price controls created new ones. The price game seems far from being successfully completed.

There has been a lively discussion about the appropriate pricing system for a labor managed economy. This discussion hardly touched the classical controversy on marginal cost versus full cost pricing. Since marginal cost pricing requires government intervention, the lack of interest in this procedure among Yugoslav economists is understandable. On a more theoretical

evel it was pointed out that allocational efficiency—as represented by marginal cost pricing—is inferior to growth efficiency—as represented by full cost pricing which makes possible the business autonomy of an enterprise (Horvat, 1964, Ch. 2).

The price debate was centered around the problem of how the price is to be formed. It started in 1950 when Kidrič opted for the "value price" (1950b). In his last writing in 1952 the late Kidrič described the value price as the one consisting of costs of production (including wages) and accumulation (gross profits) calculated as proportional to wages. These prices were actually tried out in 1953 and 1954. Kidrič compared the rate of accumulation principle with the traditional average rate of profit principle (profit proportional to capital invested is characteristic of Marx's price of production) and came to the conclusion that only the former was appropriate for a labor managed economy. In his view the average rate of profit principle "represents a contradiction to socialist planned management of the economy," and leads to "a kind of cooperative capitalism" (Kidrič, 1952 pp. 42, 46). A decade later M. Todorović—who was to become the secretary of the League of Communists—came to the opposite conclusion. He maintained that in a system of commodity production, including its socialist variety, in which fixed capital is used, prices must take the form of prices of production. Since capital is socially owned and production is planned, the use of prices of production cannot lead to the same consequences as in a *laissez faire* framework of liberal capitalism. (Todorović, 1965, pp. 60, 65, 78).

Strange as it may sound, there is no basic disagreement between Kidrič and Todorović. The difference between their views primarily reflects the difference in the degree of economic sophistication. In 1952 Kidrič's view was commonly accepted—by Todorović as well—while today hardly anybody would be prepared to support it. Todorović's theory of the specific price of production (specific because social planning is one of its basic ingredients) as an equilibrium price in the Yugoslav setting has been accepted by a certain number of economists—Z. Pjanić, V. Rakić, Maksimović (Institut, 1968)—but by no means by all. In a heated debate in Sarajevo in 1964 another group of economists—Korač, Sirotkovic, Dabčević, T. Vlaškalić—expounded the theory of "income price" (Savjetovanje, 1964). In their view the Yugoslav enterprise maximizes income in relation to suitably defined inputs. Other economists were busy inventing new types of prices: gravitational (Mesarić, 1965), normal, actual, social reproduction price (Černe, 1966, p. 233), etc. Radulović was able to describe six different price theories of this sort (1968, pp. 299–326).

Price theory is closely linked to distribution theory which we shall consider in the next chapter.

Distribution Policy

It is not conventional to talk about distribution policy. One is accustomed to speak about wages *policy* and distribution *theory*. However, as we proceed, it will become evident that in the Yugoslav setting distribution policy is also a meaningful concept.

Wages Policy: In the administrative period 1945–1952 workers were government employees classified in a certain number of salary categories according to their skills. Directorates set work norms whose overfulfillment brought an increase in pay. Managerial personnel would get premiums for the fulfillment of the government plan. The salary span was 1:3.5 (Tomic, 1968, p. 6), as compared with 1:16 before the war (Bilandžić, 1967, p. 56).

The lack of material incentives was compensated for by moral incentives such as public praise, the trophy–flag, the title of shock worker or of innovator. In the post–revolutionary atmosphere these incentives were very powerful.

After several years the lack of material incentives became a serious obstacle to efficient production. Due to post–war scarcities and to an egalitarian ideology, by 1953 the salaries of office employees in industry had been reduced by one third and of civil servants by one half relative to workers' wages and compared with the prewar levels (Berković, 1969, p. 81). Non–wage income "fringe" benefits) was higher than wage income. Since 1952 both trends have been reversed. Trade Unions advocated higher skill differentials. Economists (Bajt, 1956) urged an increase in the share of discretionary income (income after taxes and contributions left to free disposal of an enterprise) in order to increase productivity. Wachtel finds that interskill differentials increased until 1961 and then began to fall. The average income span between the highest and the lowest paid job is now 1:4 (Berković, 1969, p. 82). M. Janković estimates that wage income increased to 65 percent of total workers' income in 1956 and to 73 percent of total income in 1967 (1968, p. 159). The idea was to leave to the market the job of determining the appropriate income differentials and to stimulate efficiency by increasing the discretionary part of workers' income. The latter was also thought necessary in order to curb centralist distribution of income.

Since 1952 it has been the task of workers' councils to determine wage differentials and work incentives. The distribution of income between the enterprise and the community was settled in a very simple way. On the basis of the social plan, the expected income of the enterprise and the corresponding wage bill were determined. The difference between gross income (depreciation excluded) and wages was called accumulation and funds (AF). The ratio between AF and wages was called the rate of accumulation and funds. This rate was applied to actual gross income earned in order to derive wages. It was mentioned in section 10 that the AF rate was considered an appropriate socialist substitute for the rate of profit, and that was its theoretical justification; whatever the merit of that argument, the practical effects were good. The AF rate helped to bridge the institutional gap between complete administrative control and a relative autonomy of the enterprise. It also induced workers to economize on labor. In 1953 employment in manufacturing and mining increased by 5 percent, and labor productivity by 6.2 percent. In 1954, when the AF system was abandoned, employment increased by 13 percent and labor productivity slightly fell.

The AF rates were, of course, not uniform. The 1952 plan envisaged a rate of 19 for agriculture and another of 582 for manufacturing and mining. This difference reflected the already described goals of price policy: industrial prices were inflated in order to facilitate the collection of investment resources. However, even within manufacturing different industries had widely different rates. In industries with high rates there was no incentive to reduce costs. Since the rates could not be established very precisely, some collectives began to earn high wages. The government reacted by introducing a tax on "the surplus wage fund" (the difference between the standardized and the achieved wage bill). Since the standardized wage bill was the product of an average wage rate and the number of employed, the enterprise increased employment—often fictitiously—of less skilled workers in order to reduce the tax basis. The government reacted by differentiating taxation according to skill

categories. Enterprises countered by artificially changing the skill structure, declaring their workers to have higher skill.

The AF rates were clearly a not very refined instrument of economic policy. They were introduced in the belief that they could be standardized for all enterprises within an industry group. Soon, however, individual rates had to be prescribed for each particular enterprise. This implied direct administrative interventions which were at variance with the basic intentions of the new system. In 1954 the AF system was replaced by a system called "accounting wages," which lasted for the next three years.

Yugoslav economists had been complaining for some time that in their economic calculus enterprises do not consider capital services as a cost item (Lipovec, 1954, p. 142). That was a natural result of the fact that capital was given to enterprises free of charge. This practice was discontinued in 1954 when a capital tax of 6 percent was introduced. This tax was considered as a price for the socially owned capital and was also levied on capital invested from enterprise funds. Apart from that, the enterprise was obliged to pay normal interest rates on credits granted by the bank. Also the profit and turnover taxes were introduced, the latter becoming the chief instrument of accumulation. In this way instruments of economic policy became more varied and more flexible.

The new system implied a division of the wage fund into two components: accounting wages and wages out of profit. Accounting wages were derived by applying prescribed wage rates to skill categories taking into account actual working time. Again skills were fictitiously increased. Working time as a basis of accounting led to a disregard of work norms. The next year "wage schedules" (tarifni pravilnik) were introduced. Wage rates were determined by the social plan. Wage schedules of individual enterprises represented a kind of collective agreement between the enterprise and the Trade Union and local government (Tomić, 1968, p. 11). Differential efficiency was accounted for, and a part of profit was used as a premium for improvements of quality, reduction of costs etc. Since profit was taxed at 50 percent, enterprises tried to reduce profit by increasing wage rates and reducing norms. The government commissions for wages were unable to prevent this from happening.

In 1957 the First Congress of Workers' Councils was held. The Congress asked that the autonomy of the enterprise be widened. This primarily implied greater independence in income distribution. The division of income into wages and profit was considered inappropriate and reminiscent of wage–labor relations. In order to meet these demands, in 1958 the income distribution system was changed and a compromise reached. The wage schedules remained and were still subject to approval of local authorities and trade unions. The enterprise income was treated as one single whole and was distributed by workers' councils into wages and contributions to various funds. The difference between income and accounting wages (called minimum personal income) was progressively taxed. The wages in excess of the basic pay were also progressively taxed (Pejovich, 1966, pp. 98–99).

Progressive taxation was very much resented. And so was the outside tutorship as far as wage differentials were concerned. In 1961 both were abolished. Workers' councils became completely independent in determining wage rates and distributing income. Progressive taxation was replaced by a flat 15 percent levied on income. In 1965 even this tax was abolished.

Changes in wages policy implied drastic changes in relative factor shares. If we divide value added into gross wages (wages and taxes levied on wages) and gross

rentals (depreciation, interest, net profit and taxes levied on capital), the percentage share of the latter in manufacturing and mining varied as follows (Horvat, 1969b, p. 41):

TABLE 7

1952	10%	1961	54%
1953	11%	1963	53%
1955	74%	1964	50%
1957	77%	1965	48%
1959	67%	1966	46%
1960	62%	1967	45%

Percentage shares of gross wages represent, of course, complements to 100% of the figures quoted for rentals. In the AF system depreciation was the only capital cost. The introduction of profit and capital tax in the system of accounting wages increasedcapital cost drastically. The gradual reduction and final elimination of profit taxes, which implied a relative increase in wage tax, reduced the share of gross rental to somewhat more than one half of the value added. On these changes price changes were superimposed. The increase of food and services prices after 1960 increased nominal wages; the abolition of various subsidies at the same time and in particular after 1964 made possible a reduction in taxation which to a certain extent offset the effect of wage increases. The next effect was to lower the share of gross rental below 50 percent. The adaptation of an enterprise to these changes required an extraordinary effort on the part of the management. But enterprises did react. Simultaneously with increased capital charges the capital coefficient (the ratio of gross fixed capital to gross material product) in manufacturing and mining fell from 3.6 in 1955 to 2.5 in 1964 (Horvat, 1969b, p. 51). If enterprises are market oriented and if the production function is linear homogeneous (which proved to be an acceptable approximation), the elasticity of output with respect to capital in the last decade must lie somewhere in the region 0.45 to 0.62. The actual elasticity coefficient turns out to be 0.48. This is taken as one indication that the economy is following market rules (Horvat, 1969b, p. 42).

While wage systems with wage schedules and progressive taxation were applied, real wages lagged behind productivity increases and producer prices were stable. From 1958 on real wages began to increase faster than labor productivity, and the discrepancy between the two series was widened particularly in the cyclical trough in 1961/62 and after 1964 (Popov, 1968, p. 627). The peculiar movements of prices that followed were considered in the section on Price Policy. Another peculiarity was established by Wachtel: interindustry wage differentials continued to increase, and interindustry wage structure appeared as a function of average productivity which explained 80 percent of the variance (Wachtel, 1969, pp. 151, 175). Popov found a high correlation between the rate of growth of industrial output and the productivity of labor ($r=0.86$) (Popov, 1968, p. 622). If all these bits of information are put together, the following interpretation begins to emerge.

Trade Unions announced the principle: wages should increase proportionally to the productivity. The principle was widely accepted, and it is a sound principle when applied to the economy as a whole. If applied to individual enterprises, it generates great trouble. In a rapidly growing economy various industries expand at widely different rates (petroleum industry at 19.2 percent, tobacco industry at 5.1 percent per annum in the period 1952–1966). Thus rates of growth of labor productivity are bound to differ very much (11.7 percent and 1.2 percent respectively). Thus wages must differ and differentials must increase in time (money wage rates

increased 12.8 times in the petroleum industry and 8.3 times in the tobacco industry in 1952–1966) (Popov, 1968, p. 630). Kovač found that in 1966 wage rates for the same category of skill in the highest paid and the lowest paid industry group were related as 2:1 (1968, pp. 130–33). All this is, of course, in flagrant contradiction to the principle of distribution according to work. That is why Bajt remarked that the principle of remuneration according to productivity actually denied the principle of remuneration according to work performed (1967b, p. 363). Deviations of productivity income from labor income have been analyzed by the present author. They represent (after deductions for other facto costs) a form of rent which I call the rent of technological progress (Horvat, 1962b). The faster the rate of growth, the more important this rent becomes.

Rašković (1967b, p. 230) and others suggested that the principle of distribution according to work be replaced by a more appropriate principle "according to the results of work." It is not the process of work as such but its results that have to be rewarded. Rašković noted that grossly imperfect markets in Yugoslavia meant exploitation of one group of collectives by another, more privileged, group (1967b, p. 218).

The meaning of the principle, "according to the results of work" has been stretched by Šefer in a rather curious fashion. Šefer notes that in developed capitalist countries free market wage determination has been replaced more and more by a policy of "equal pay for equal work." He feels that such a policy is inapplicable in Yugoslavia because workers bear business risks, i.e., they share in both profits and losses. Work cannot be remunerated automatically; it has to be socially recognized, which happens at the market where the exchange determines the result of work. The principle "equal pay for equal work" could be implemented only in a system of state ownership and state management of the economy (Šefer, 1968b, pp. 74–75). Thus Šefer, Korač and a certain number of others in fact argue that the principle, considered Marxian, can be implemented in a capitalist and étatist setting, but not in a self-government system. The fallacies of this *laissez faire* reasoning are obvious: market imperfection provides no criteria for the social recognition of somebody's work; the redistributive effects of market imperfections can be eliminated also by means other than the étatiste ones.

Other Issues: On income differentials due to technological and other rents, differentials due to variable entrepreneurial abilities of various working collectives are superimposed. Šefer quotes data for Belgrade enterprises in 1967 when the same jobs in various enterprises were paid rates as different as 1:3 or 1:4 (1968a, p. 434). It is clear that such extreme differences generate enormous inflationary pressure. There is also an additional consequence. Capital intensive enterprises are able to improve their personal income position by distributing a part of profit in wages. That is why wage rates are positively correlated with capital intensity. Yet, if profits tend to be reduced, enterprises become more and more dependent on outside sources for financing their investment. This generates new difficulties which we will consider in the section on Banking and Monetary Policy.

Apart from technological rent, the classical forms of rent were both discussed in the literature and applied in practice (Bakarić, 1950; Horvat, 1953). Agricultural rent is absorbed, in principle, through taxation according to cadastral revenue. Mining rent represented a separate item of income of mines and crude oil producers for several years. However, it was determined in a rather arbitrary way and gen-

erated regional differences. Consequently, it was resented by the enterprises and was eventually abolished. All urban land belongs to communes and urban rent is used to finance communal investment.

J. Dirlam, an American student of Yugoslav economic affairs, points out that the Yugoslav system can be viewed as one in which labor employs capital, instead of a system in which capital employs labor as is the case under capitalism. The social ownership of capital requires a somewhat different approach to capital charges in the labor-managed enterprise as compared with its capitalist counterpart. The floor and not the ceiling is set for depreciation rates. Profits need not be taxed and instead payroll taxes are suggested (Institut, 1968b). A tax on capital is primarily an instrument for allocating resources and not necessarily a device for collecting revenue for the government. The revenue from capital taxation has been used by the government to finance major investment projects and also to finance the Fund for underdeveloped regions. Resentment against these redistributive activities of the Federal Government has been growing, and recently a political decision has been made to abolish the capital tax. Many economists disagree with this decision. Some argue that the abolition of capital tax, which represents the price for the use of social capital, will initiate a transformation of social ownership into collective ownership. D. Gorupić and J. Perišin argue that the price of a product should contain an element of growth (1965, p. 124). This is to be achieved if accumulation is determined by the social plan in the form of interest on capital used. But this money must not be expropriated by the state; it ought to remain in the enterprise earmarked for investment. Thus this internal interest is to be treated in the same way as depreciation. In order to cope with business fluctuations, minimal depreciation cum accumulation must be determined in a cumulative fashion (Gorupić, 1968, pp. 12, 13). Lavrač maintains that the accumulation-protecting interest rate may be differentiated according to industries and regions (1968). S. Popović suggests that the compensation for the use of social capital will provide the bulk of development resources. After all factors of production, except labor, are paid their shares, the remaining net income is to be distributed among workers. Additional accumulation can be derived only from this private income, which means that workers remain owners of that part of capital (S. Popović, 1968). Similar is the position of Černe who maintains that the participation of workers with their own means in the development of the enterprise—which implies receiving adequate interest or dividends—would stimulate rational behavior of workers and management bodies (1967a, p. 21). On the other hand, Samardžija argues that this is both economically irrelevant and socially dangerous. Contemporary shareholders participate in the profits of their corporations with only small percentages that accrue to dividends. And attempts to make workers co-owners must end in the establishment of a separate group of owners of means of production within the society (Samardžija, 1968, pp. 145, 303).

We have thus reached the point at which the general principles of an adequate distribution policy may be discussed. There seems to be considerable agreement on two issues. (1) As great a part of income generated as possible should remain under the direct control of the working collective. (2) Only labor income should be distributed in wages. These two principles imply a sharp division of income into two components: labor income appropriated by individual workers and nonlabor income belonging to the society but remaining under the control of the working collective and

used exclusively for investment purposes.

In order to be able to divide net income into its labor and nonlabor parts, we need a theory of factors of production. In this respect Bajt follows the traditional approach and defines factors of production as sources of productive services. He enumerates five such sources: labor, entrepreneurship, invention, land and capital (1967b, p. 351). The first three generate labor income, although normally a small proportion of income from inventions is appropriated by inventors. This theory leads Bajt into difficulties when he has to explain monopoly income. He then argues that in a market economy monopoly participates in income; monopoly does not add to output but only adds to income of all factors (1967b, p. 357).

In order to avoid the shortcomings of the traditional theory, the present author defines factors of production as types of forces that influence the generation of output. Factors have to be priced in such a way as to lead to an optimal allocation of resources. The latter means achieving maximum output from given resources or minimum input of resources for a given output. There are four factors: labor, entrepreneurship, capital and monopoly. The first two generate labor income (wages and profit), the latter two generate non-labor income (interest and rent). Creative work and organizational work as well as routine work generate labor income. The income due to the activities of the work collective as a whole represents entrepreneurial income. Capital services are priced in the usual way and have already been discussed. A few more words need to be said about the morphology of rent. Rent is the price of monopoly in the sense that it represents the surplus over the minimum supply price of resources. Land rent appears in three forms described by Marx (differential rent I and II and absolute rent), then there is mining rent and a somewhat special urban rent. The rent of technological progress—due to the fact that certain industries expand faster and enjoy economies of scale effects, or participate more in general technological advance, or both—has already been described. Bajt adds the rent from market monopoly, which he describes as a situation when the selling prices are above normal and the buying prices are below normal (Bajt, 1962, p. 93). After land, natural resources, technology and market monopolies are accounted for, the remaining part is a monopoly in the narrower sense. Except for the last, the prices of the other monopoly factors may be in principle determined either by the market mechanism (land and mines) or by economic analysis (technology and market). As far as the latter is concerned, progressive taxation may in practice prove a more efficient procedure. If taxes are designed in such a way as to be generally considered as just, they will not affect the supply of resources and this is how in fact we defined rent (Horvat, 1964, ch. 3, 4, 6).

Actual business practice and legislative measures do not quite follow the principles discussed above. The productivity-wage practice leads to an appropriation of a considerable part of non-labor income. The same consequences follow from the facts that mining rent is included in undifferentiated income and that there is no progressive taxation. In 1968 the new law on the distribution of income in the enterprises included income from capital invested in other enterprises in the undifferentiated income of the collective-investor. P. Jurković promptly called that a rather dubious theoretical solution (1969, p. 50). In general, the distribution of income according to the work performed is still a goal to be reached.

Foreign Trade Policy

Background: The pre-war trade struc-

ture was rather simple. Food and other agricultural products represented about one half of total Yugoslav exports. One fifth of exports consisted of wood and almost an additional fifth of non-ferrous ores and metals (Dobrinčić *et al.*, 1951, p. 408; Fabinc *et al*; 1968a, p. 144). Thus close to ninety percent of export earnings were provided by these three sectors producing raw materials and semi-manufactured goods. Immediately after the war the development strategy consisted in (1) expanding the exploitation of natural resources in these three sectors and (2) in using the export proceeds to finance imports of equipment and other producer goods. It was also expected that (3) the Soviet Union would provide great help in speeding up economic development. The second part of the program was carried out successfully, the share of consumer goods in imports was reduced from 22 percent before the war to only 11 percent in the period 1947–1951 (Čehovin, 1960, p. 59). The first and the third parts encountered unexpected difficulties.

Due to a decline of per capita agricultural production and rapid industrialization, agricultural export surpluses were reduced and so was the total volume of exports. It soon became fashionable to explore the question whether Yugoslavia was not becoming a permanent net importer of agricultural products (Srdar, 1953). The nationalization of foreign property imposed a new burden on the balance of payments. On the other hand, reparations for war damages were to a large extent left unpaid, in particular by Western Germany and Hungary. Immediately after the war about 75 percent of foreign trade was conducted with the Soviet Union and her East European allies. In 1947–1948 the trade shares with these countries were stabilized around 50 percent in exports and 42 percent in imports. In the middle of 1948 the ominous Resolution of the Cominform meant the end of good relations. By 1949 the Soviet group reduced the trade to one-third and in 1950 it was cancelled altogether. The Soviet Union and her allies applied a total boycott to all relations with Yugoslavia.

Thus the country was cut off from the East completely. It was separated from the West as well, as it did not enjoy the facilities mutually provided by western countries to each other. It was not included in the Marshall Plan; it remained outside GATT. In short it was isolated in a hostile world. The five-year industrialization plan—imbued with so many hopes—had only been initiated, when suddenly the contracts were broken, and supplies of equipment and materials ceased to arrive. Trade was declining:

TABLE 8

	1948	1949	1950	1951	1952	1953	1954	1955	1956	1965
Exports	100	79	74	64	87	80	102	99	122	328
Imports	100	95	86	114	115	106	103	130	142	311

Sources: *Jugoslavija 1945–1964*, p. 77: *SGJ*-1959, p. 121.

Foreign exchange reserves dropped from 43 percent of the value of imports in 1937 to 12 percent in 1948 and to 4 percent in 1952 (Mrkušic, 1963, p. 186). Personal consumption was declining. Defense expenditures amounted to twenty percent of national income. Two severe droughts, on in 1950 and the other in 1952, proved unexpected allies of the Cominform and reduced agricultural output to 25 percent below the

pre-war average. The situation looked hopeless. That is why Stalin expected surrender.

Yet this nation was not accustomed to surrender; it was more at home in fighting back. And it did so, for the first two years struggling practically alone. Investment plans were changed, trade was channelled towards the West, even the economic system was changed. From 1951 on, foreign economic aid began to flow, mostly from the United States. It consisted primarily of food, raw materials and military supplies. The aid amounted to 38 percent of total imports in 1951, and over the next decade was gradually reduced to zero.

The crisis was soon overcome and the economy entered a period of unprecedented growth. The effects of the heavy capital investment of the First Five-Year Plan began to materialize in rapid expansion of industrial output. The new agricultural policy soon generated phenomenal growth of agricultural output. Exports were catching up with imports. In 1954 the first trade contacts were established with the East European countries. After the conciliatory visit of Premier Khruschev to Belgrade in 1955, normal trade relations were established and so a precious outlet for increasing exports was found (Obradović, 1962, p. 40). In the decade that followed, exports increased 3.3 times, i.e., at a rate twice as high as in the world as a whole.

These developments were too good to last long. In 1957 the Common Market was born in Rome. Two years later EFTA was created in Stockholm. Practically all West European countries became members of the one or the other trading group. East European countries belonged to COMECON, created in 1949, but actually operating since 1954. Yugoslavia found herself isolated again. At first it did not matter too much. But gradually intrazonal trade in all three areas began to increase rapidly and to depress trade with third parties. This was true in particular for the Common Market, the most important trading partner of Yugoslavia. Common Market countries account for 30 percent of Yugoslav exports, 38 percent of imports and two-thirds of financial transactions. What makes this trade so vulnerable is the fact that between one third and half of Yugoslav exports to Common Market countries consists of agricultural products. Regular and variable import tariffs in the Common Market amount on the average to 50 percent of the Yugoslav export prices, for beef even to 60-70 percent, which clearly cannot encourage exports. Variable protection rates, when first announced to GATT, were said to be an exceptional instrument, the customs tariff remaining the basic one. In fact, however, variable rates amount to 2.5 times the regular tariff, they are changed daily, weekly or quarterly and represent a permanent instrument of total protection (Žiberna, 1969; Mitić, 1969).

Yugoslavia reacted to the new situation by trying to increase her trade with the developing countries. This attempt met with a limited success. Imports from developing countries increased to a maximum of 14.1 percent of Yugoslav imports in 1964 and there has been a permanent balance of payments surplus with these countries (Pelicon, 1968). Next, close relations were established with GATT. At first an observer, Yugoslavia became an associated member of GATT in 1959 when she also enacted the Customs Law. In 1961 a temporary customs tariff was produced and next year Yugoslavia became a temporary member of GATT. In 1965 a new, permanent customs tariff was enacted, and a year later full membership was granted by GATT.

COMECON was also approached. Its members absorb almost one third of Yugoslav trade. In 1964 Yugoslavia be-

came an observer in COMECON. With the Common Market special agreements are negotiated.

India and the United Arab Republic account for one third of Yugoslav trade with developing countries. In 1966 the heads of the three countries initiated a scheme which became known as Tripartite Co-operation. The agreement, ratified in 1968, comprised 500 products to which preferential rates of 50 percent became applicable, and envisaged also industrial co-operation. It was also suggested—this time by economists and not be politicians (Bilandžić, 1967, p. 33)—that a Danubian trading area be formed. If that had proved possible, it was hoped that the area could have been extended North and South. The occupation of Czechoslovakia rendered that idea utopian for the time being.

Attempts to develop economic relations with as many countries as possible and the foreign policy of an uncommitted nation enabled Yugoslavia to establish trade with 120 countries. Trade is not only geographically dispersed, it is also diversified in terms of products exchanged. As a result a theory of "capillary trade" emerged. V. Pertot argues that small quantities reduce marketing difficulties, and S. Obradović adds that highly diversified trade reduces risks of business fluctuations. Empirical research lends some support to this hypothesis. P. Mihajlović finds that the concentrated pre-war export was very much dependent on external business fluctuations, while no such dependence appears to exist after the war (Mihajlović and Tanović, 1959, p. 77). Capillary trade also has its drawbacks. Obradović points out that it increases marketing costs and quotes approvingly Bičanić, who maintains that export concentration is a precondition for a permanent export position on the world market (Obradović, 1962).

Fast growth after 1955 led to profound structural changes. The share of exports of commodities and services in social product increased from about 13 to about 20 percent. The Yugoslav share in world trade doubled, but being still less than one percent, provides a justification for the capillarity theory. The share of those three traditional natural resource sectors in exports has been reduced from 90 to 50 percent (Fabinc *et al.*, 1968a, p. 144). Raw materials and manufactured goods changed their places in the structure of exports (Guzina, 1950:6 in 1939 to 13:50 in 1968). The once self-sufficient peasant economy is now only a matter of historical interest. It has been replaced by a relatively open industrialized economy participating actively in development of the world market.

Prologue: Rigid central planning in the period 1945–1951 implied a state monopoly in foreign trade. The domestic market was completely cut off from the outside world. The rate of exchange was just an accounting device without economic meaning. Export and import trade were conducted at prescribed domestic prices. The Fund for Price Equalization, created in 1946, compensated exporters for the differences between the domestic and export prices. Each transaction implied a separate foreign exchange rate. That was consistent with the principle of profitability at all cost applied in the home market. Exporters were obliged to surrender their foreign exchange proceeds to the National Bank which, in turn, supplied importers with what they needed. Foreign trade enterprises acted as agents for the Ministry of Foreign Trade and were obliged to implement import and export plans. Plans were defined in physical terms and so traders were not interested in prices and other trading conditions. The system was simple and consistent, but not very efficient. Yet, in the turbulent post-war years it did the job it was designed for.

The most important event in those years was the Cominform economic boycott. At

that time details about operations of mixed Soviet–Yugoslav companies became publicly known and stirred great indignation. A certain number of these companies were created with a proclaimed aim of helping to develop the country. Capital was invested in even shares, profit was divided evenly, the Russians appointed their own people as general managers, insisted on preferential treatment and objected to Yugoslav financial control. All this reminded people too much of their pre-war experience with foreign capital and mixed companies were gradually liquidated. But the problem was more complex than that; economic relations among socialist countries were at stake.

In an interesting 1949 article M. Popović, then a member of the government and now the President of the Federal Assembly, explained the position that had been taken (1949). If a less developed and more developed country meet in the world market, they will exchange commodities with different labor contents. The more productive country will get back more labor than it gives away. This implies exploitation. Further, if in mixed companies profit is divided according to capital invested, a principle of distribution alien to socialism is introduced and as a result exploitation appears in yet another form. "According to socialist principles"—said Popović—"the entire surplus value, i.e., the entire profit obtained by the society after it had sold the commodity in the world market, belongs to the proletariat which has created that value . . ." (1949, p. 108).

To such theories, and not quite unexpectedly, Russian negotiators reacted rather laconically: "Torgovlja—torgovlja, a družba—družba" (trade is trade and friendship is friendship). But for Yugoslavia, then a year or two after the Revolution, socialism meant immensely more than trade; to put the two on an equal footing was profoundly shocking. Economic relations among socialist countries were seen as similar to the relations of the various regions within one country. Developed socialist countries had an obligation to grant aid to the less developed ones in order to speed up their growth and enable them to reach the same level of development in the shortest possible time (Obradović, 1962, p. 39; M. Popović, 1949, p. 70).

These were not abstract ideas; they were applied in relation to Albania. Yugoslav and Albanian partisans fought together during the war and relations between the two countries were very close. As a more developed country, Yugoslavia sent experts and material supplies to Albania. Tariffs were abolished and monetary units were given the same nominal value. Attempts to design a single system of prices failed because productivity differences between the two countries were too great. But they then continued to trade at their internal prices which meant that Albania exported at Albanian prices and imported at Yugoslav prices (the latter were somewhat lower than the Albanian on the average). In this substitution of world market prices by respective domestic prices. Popović saw the elimination of the exploitation characteristic of the world market mechanism (M. Popović, 1949, p. 128). In fact, however, this conclusion does not necessarily follow. To find out whether and how much Albania gained, one would have to calculate the entire trade in Albanian, Yugoslav and world prices and compare the value aggregates. And in order to make exchange equivalent in labor terms one would have to apply input-output analysis. Another policy measure had much more obvious implications. Albania was granted interest-free loans for an unspecified length of time. This was an early anticipation of the now familiar aid programs for underdeveloped countries.

Bulgaria was another country with which

Yugoslavia expected to eliminate tariffs and possibly even form a confederation. Yugoslavia waived Bulgarian reparations obligations for war damages, and after the Bled agreement in 1947, hopes went high in both countries. A few months later Stalin launched his attack, and soon all achievements were forfeited, all hopes buried. Former friends became enemies.

The Cominform economic boycott and the need to finance the Five-Year Plan compelled Yugoslavia to establish contacts with the world capital market. Ideological reasons and unpleasant experience with Western capital before the war and with Soviet capital afterwards made joint stock companies and mixed companies an undesirable form of import of foreign capital. Loans remained the only available alternative. But loans may also affect the economic and political independence of the country unfavorably. In order to prevent this from happening, V. Guzina suggested, in a paper representing the common opinion of the time, that foreign trade be conducted according to the economic plan, and a specified volume and structure of exports be secured (1950, p. 71). Guzina also held that autarchy was both impossible and undesirable, and favored development of an open but controlled socialist economy. These ideas were characteristic of foreign trade policy in the next decade.

Three Steps Towards Free Trade

By the middle of 1951 the new economic thinking reached the sector of foreign trade. As usual, market experimentation began with agricultural products. Exporters of certain agricultural commodities were allowed to sell their foreign exchange proceeds at a price which was obtained by multiplying the official rate by the factor 7. This foreshadowed the new official rate determined on January 1, 1952 at 1$=300 din (the old rate was 1$=50 din). Exporters were granted a retention quota of 50 percent with which they could finance imports of their own choice and sell imported commodities at free prices.

The transition from complete state monopoly to a system of free trade was not a simple affair. Various alternatives were discussed. In an important article early in 1952, D. Avramović, now a staff member of the World Bank, argued that a fixed exchange rate and, in particular, its exclusive use, cannot be practiced in a socialist economy. In order to secure the minimum volume and the necessary structure of exports and imports consistent with production and investment targets, the fixed exchange rate should be replaced by either physical allocation of goods or a system of multiple exchange rates. The latter is more consistent with a socialist market economy. Since foreign prices constantly fluctuate and since a full employment high rate of growth economy needs stability, there ought to be an Equalization Fund to absorb violent fluctuations. Thus, not only is there a need for multiple exchange rates, but these rates should also fluctuate. The capitalist principle of a fixed exchange rate *cum* business fluctuations must be replaced by a socialist principle of multiple fluctuating exchange rates *cum* economic stability and growth (Avramović, 1952).

Most of these ideas were soon tried out. In July of the same year the system of 17 price equalization coefficients was set in operation. Coefficients, applied to export prices calculated at the official exchange rate, ranged from 0.8 (for exports of agricultural products) to 4.0. Low coefficients were applied to imports of equipment and raw materials in order to keep their prices low.

A high degree of liberalization was envisaged in foreign trade, but in comparison to the liberalization of the home market, the liberalization of the foreign trade system proved to be a much tougher job. First of all—and again in contrast to the

home market—the price of foreign exchange was set too low. Already in 1951 the actual average export exchange rate was 354 dinars for one dollar, and in 1952 it increased to 585 dinars which was almost twice as much as the official rate. The average import exchange rate was lagging behind appreciably (1$=440 din). Cheap imports exerted pressure on the balance of payments. The foreign exchange reserve of 4 percent of imports made economic interventions impossible. No wonder that the newly created foreign exchange market, DOM (Foreign Exchange Accounting Place), did not work. At first, exporters were obliged to sell only 55 percent of their foreign exchange to the Bank; the remaining 45 percent, representing their retention quota, could be used for imports of their own choice or sold to importers at the DOM. Already in October the retention quota was lowered to 20 percent, and that meant the death sentence for DOM. In the next year DOM rates soared to a level 6.8 times higher than the official rate. Average actual exchange rates went up as well.

In 1954 a series of desperate attempts was made to save the system. The accounting exchange rate was increased to 632 dinars for a dollar. Coefficients were revised and applied to DOM rates, and not to the official rate. A steep tax on the gains at DOM was introduced. A number of other complicated procedures were applied. DOM rates were brought close to the new accounting rate, which the authorities aimed for. Yet importers of raw materials could not compete any more at DOM for foreign exchange and so separate sales were organized for them. This reduced the amount of available free foreign exchange to something like one percent of the demand. The retention quota was reduced to only one percent. Prices of foreign exchange soared and by 1960 reached a level 12.3 times as high as the official rate. The National Bank replaced exporters as the only seller of foreign exchange (Mrkušić, 1963, pp. 301–315).

The first free trade attempt failed because the initial price for foreign exchange was set too low, initial reserves were too small, the share of the free market in foreign exchange supply was too small and disparities between home and foreign prices too great. It would have been rather difficult to find elsewhere in the world such relative prices, remarked V. Meichsner, as existed in Yugoslavia in 1955: one typewriter ribbon (2.800 din) equals a pair of shoes equals two yards of woolen fabric equals one third of an average employee's salary equals two-day full board in a first class hotel in a tourist resort equals 56 haircuts equals the monthly rent of a five-room apartment (Majhsner, 1956, p. 193). At that time three different foreign exchange regimes coexisted: the official rate, the regular and the separate DOM rates. Meichsner suggested that the number of coefficients be gradually reduced to only two, one for industrial and one for agricultural products. In 1957 M. Frković calculated deviations of actual exchange rates of various product groups from the average actual rate of 1$=779 din. It turned out that industrial exports and food, equipment and invisible imports were subsidized at rates between 21 and 35 percent, that there were export taxes between 16 and 21 percent for agricultural, wood and invisible exports and a protection rate of 105 percent for consumer goods imports (Frković, 1957).

By 1960 it had become clear that the foreign trade system needed a thorough revision. D. Čehovin evaluated the situation in three points. Enterprises were stimulated to press for an increase in coefficients, not to compete in the world market. Coefficients had ceased to be passive equalization instruments and were in fact transformed into active devices for increasing price disparities. Profitability calculations were made practically impossible

(Čehovin, 1960, p. 125). Mrkušić noted that in an economy where exports are price elastic and imports are not, there will be a constant tendency for export exchange rates to move away from the import ones. That required physical restrictions on imports (Mrkušić, 1963, p. 297). Both did happen. Higher export exchange rates were bound to produce inflationary pressure—via the money supply—as Avramović had already warned (1952, p. 24).

The recession that started in 1960 made things worse and stimulated the authorities to undertake a reform in 1961. This time an ample supply of foreign exchange was secured by foreign loans. But the other two mistakes of the 1952 reform were committed again: the new accounting rate was set too low (750 dinars for one dollar); the actual export rate in 1960 was 981 din and in 1961 went up to 1021 dinars (O. Kovač, 1966) and price disparities were corrected in only a few cases.

The strategy of the reform can be described as follows. Multiple exchange rates were abandoned and coefficients were replaced by a customs tariff. Instead of exchange rates varying between din 500 and din 1200 for a dollar, there was to be a single 750 rate with no protection for agriculture and lumbering, with 10–40 percent protection for consumer goods and 17–60 percent protection for equipment and other industrial products. Export was free and was supported by premiums and tax reductions. Exporters were supposed to sell foreign exchange to the National Bank but in most cases could buy back 7 percent of the amount sold for their own needs. About one fifth of imports was liberalized, and for the rest commodity quotas or foreign exchange allocations applied.

The deficiencies of such a strategy soon became apparent. Exports were retarded, imports accelerated. In order to keep the balance of payments deficit under control, import restrictions were multiplied and in 1964 the tariff protection was increased from 20 to 23 percent. Exports were stimulated by making foreign exchange allocation conditional upon export sales. Export premiums and tax reductions were rapidly expanding. Soon the old system of multiple exchange rates reappeared with all its inefficiencies (Institut, 1964).

The situation was worsened by the fact that about one half of Yugoslav foreign trade was oriented towards clearing currency countries, most of it towards COMECON. Both import and export flows with the COMECON countries are much more unstable than with the convertible market (Madžar, 1968). Both import and export prices on the COMECON market are higher than on the world market. Besides, it is easy to export to this market but difficult to import from it and vice versa for the convertible currency market. As there was one single exchange rate for both markets, the consequences should be obvious. Importers were oriented towards convertible currency countries, exporters towards clearing currency markets. The balance of payments deficit with the former increased rapidly, while there was an unabsorbed surplus on the trading account with the latter. A boom in 1964 produced unbearable pressure on the balance of payments. In the same year the cycle was reversed. The recession helped to induce the authorities to undertake another reform in 1965.

This time the structure of domestic prices was radically readjusted as explained in section 10. The actual export rate of exchange in 1964 was 1050 din; it was expected to increase in 1965 to 1200 din and the new official rate was determined at 1$=1250 din. Thus two fatal mistakes of two preceding reforms were avoided.

An additional element in the strategy consisted in the lowering of tariff protection from 23.3 percent to 10.5 percent with the traditional differentiation of rates from 5 percent for primary commodities to 21 percent for consumer goods (Domandžič, 1966). The necessary supply of foreign exchange was secured through the cooperation of the International Monetary Fund (IMF).

The ambitions of the reform were great D. Anakioski, one of the directors in the Federal Planning Bureau, describes the objectives of the reform as follows. The Yugoslav economy was to be integrated into the world market. Trade was to be gradually liberalized and the dinar made convertible. Exports were to rise relative to imports which would permit building up substantial foreign exchange reserves. The balance of payments deficit was to be eliminated (Anakiovski, 1969).

The new foreign trade regime became operative in 1967. About one quarter of imports was liberalized and retention quotas remained in most cases at 7 percent. For the rest there was a complicated system of inducements and restrictions; in order to achieve a proper regional distribution of trade, a category of imports from the convertible area was made conditional upon the purchase of a specified amount of clearing currency (Saveana, 1966). Export premiums and tax subsidies were abolished. Tight money policy was to keep prices stable, reduce internal demand and compel enterprises to export.

Once again the new regime failed to produce the results expected. After an initial burst of exports and a contraction of imports which in 1965 produced a small balance of payments surplus, imports began to expand faster than exports. Internal demand was checked, but so were exports. A balance of payments deficit reappeared and was increasing. Unpleasant clearing currency surpluses were cumulated. Import restrictions were multiplied. Export inducements were reintroduced. Differential exchange rates were back. The dinar was stable on the tourist market—dinar notes could be bought at rates close to the official one at all foreign exchanges—but a quiet devaluation was proceeding under the surface. None of the objectives quoted by Anakioski was achieved.

The ways in which the free trade reforms were carried out did not indicate an impressive professional competence. But in this respect Yugoslavia is not unique in the present world. The most popular method of policy making seems to be the method of trial and error. It has its drawbacks but, if applied with sufficient persistance, it also produces useful results.

So far I have been examining deficiencies. Let me now briefly evaluate the results. Since 1952 the span between extreme actual (resulting from actual revenues of exporters and actual payments of importers) exchange rates has been considerably narrowed. Actual exchange rates have become considerably more stable. The positive difference between the actual export and the actual import exchange rates of 303 din. in 1955 was transformed into a negative difference of 100 din in 1967. Government interventions in foreign trade operations have been reduced in every respect. About one fifth to one fourth of imports is firmly and completely liberalized either directly or via retention quotas and other arrangements. The tourist dinar is a stable and convertible currency. The stage is set for the last—if there is such a thing in economics—assault on free trade and convertibility.

The What-to-do-Next Controversy

The misfortunes of the third reform were not entirely unexpected. Mrkušić, A. Čičin-Šain and other economists evaluated various government objectives as unattainable given the policy pursued.

Soon a lively discussion developed focussing on three themes: protection, the nature of exchange rates, and convertibility.

I. Fabinc argued that every protection policy ought to be associated with a development program. Developing countries encounter serious bottlenecks in output capacities and shortages in material and financial means. Therefore, unlike developed countries whose protection policy aims at changing the structure of prices and incomes, developing countries must have a protection policy oriented towards changes in the structure of production. The main objective of tariff policy is to protect national production by producing a desirable differentiation of internal prices as compared with prices on the world market. There are, however, three important tasks which a tariff policy cannot perform. It cannot regulate the volume of imports, it cannot achieve the desirable structure of imports and it cannot regulate a regional distribution of trade (Fabinc, 1963, 1968b). One has to find other devices to do these three jobs.

Evidently, administrative interventions of the government are one possible alternative. It is, however, not acceptable as a dominant alternative in the Yugoslav setting. Next, a proper exchange rate system could do at least part of the job. This system could be based on one single rate, or on multiple rates, and the rate or rates could be pegged or be fluctuating. Out of these elements four main combinations and a number of variations may be formed. On the one extreme there will be a single pegged rate and on the other fluctuating multiple rates.

In the debate the Institute of Foreign Trade noticed an inconsistency in the traditional approach. The policy of a single rate usually imposes the elimination of multiple rates on the export side, while on the import side they are retained in the form of a customs tariff. In fact, however, the economic justification for multiple rates is the same for both components of foreign trade (Institut, 1964, p. 75). Mrkušić and O. Kovač of the IES suggested that the pegged rate be made flexible by the application of exchange rate ingredients such as tax reductions, preferential transportation rates and the like. But they find direct export subsidies unacceptable, presumably because they fear a proliferation of arbitrary government interventions (Bilandžič, 1967, p. 34). As far as the import side is concerned, Fabinc noted that fixed customs rates do not prevent their flexible application (by an appropriate definition of the customs value or by introducing point clauses) (Fabinc, 1963, p. 38). Other devices—such as a customs registration tax—are available as well. Thus even if the single fixed exchange rate is chosen as a basis for the system, the prevailing expert opinion favors making it flexible in both senses: it ought to be changeable in time and differentiable with respect to the fixed standard. The justification for this approach had already been provided by Avramović in the cited paper of 1952: a planned economy cannot tolerate that outside economic conditions and fluctuations be automatically transmitted to the internal market. This was now reiterated by U. Dujšin, who advocated not only flexible, but also fluctuating rates (Dujšin, 1968, p. 593). Mrkušić pointed out that if one wanted to keep the balance of payments in equilibrium either the exchange rate or internal prices will have to be continually adjusted. Since internal stability is obviously the first priority, the flexibility of the external value of money follows as a natural consequence (Mrkušić, 1967).

The government chose to base its policy on the pegged rate. This decision now came under attack. A pegged rate implied government interventions, which were

resented. Fluctuating rates involved risks of instability, which the government was not willing to assume. Čičin–Šain thought that these risks could not be so great, that fluctuating rates required much smaller reserves and much less stringent conditions in terms of financial discipline, organization of the market etc. (Čičin–Šain, 1968b). A few years earlier, G. Macesich, an American economist of Yugoslav extraction, also argued in favor of fluctuating rates. He believed that "such a system would serve to integrate the country's economy more effectively with the world's economy by quickly indicating to planners when mistakes in the planning have been made. The correction of mistakes would not have to depend on intermittent changes in rigid official exchange rates" (Macesich, 1964, p. 202).

On the other hand Mrkušić argued that fluctuating exchange rates would generate speculation and would be destabilizing. He cited the Canadian twelve-year experimentation with fluctuating rates which he claimed ended with trade restrictions for about one half of imports (Mrkušić, 1969). Čičin–Šain suggested that speculation could be avoided if enterprises were obliged to sell foreign exchange as soon as it was earned. Capital movements would clearly require separate control.

Fluctuating exchange rates implied the existence of a foreign exchange market. The government feared that this might mean repeating the failures of the DOM. On the other hand enterprises and business chambers were pressing for higher retention quotas. The prevailing export opinion seemed to be in favor of the market, even if not for all currencies. Since the country ran a chronic surplus on its trading account with the clearing area as a whole and with most individual clearing currency countries, it seemed advisable to start market operations with these currencies (Bilandžič, 1967, p. 34). That would mean fluctuating rates for about one half of the foreign exchange proceeds. The next phase might be trading in convertible currencies, and finally a proclamation of the external convertibility of the dinar.

Čičin-Šain examined the pros and cons of approaching full convertibility via external convertibility, i.e., by satisfying Article VIII of the IMF agreement, or via internal liberalization. In favor of the former, he advanced the following three reasons: (1) the dinar might become a reserve currency, which would mean an interest-free credit for Yugoslav imports; (2) clearing countries might find it advisable to liquidate their clearing deficits in order to accumulate convertible dinar balances and (3) the financial prestige of Yugoslavia would increase. He felt, however, that these reasons were not particularly convincing. Even if fully convertible, the dinar would probably not be held as a reserve currency in any substantial amount, and in so far as clearing deficits were structural, they would not be remedied by financial devices. On the other hand, external convertibility would require substantial reserves and is the more difficult to achieve the higher the degree of internal liberalization (Čičin–Šain, 1967, 1968a). Liberalization would result in lower inventories—inventories are notoriously high in the Yugoslav economy—which would mean a considerable saving in foreign exchange and in working capital.

Later in the debate professional opinion swung in the direction of external convertibility. Mrkušić argued that in fact Yugoslavia maintained external convertibility with the convertible currency countries. If Yugoslav traders pay foreign exporters in their own currency, this is the same as if they paid in an externally convertible dinar. The official proclamation of external convertibility would lead

to greater financial discipline, greater influence of the world market on internal costs of production and also to some foreign exchange economizing because foreign exporters would not insist on converting dinar balances immediately into their own currency (Mrkušić, in press). The Economic Institute in Zagreb pointed out that external convertibility would facilitate multilateralization of trade with the COMECON countries (Fabinc *et al.*, 1968a, p. 191).

As already noted, Yugoslavia belongs to neither of the trade areas in Europe and is politically uncommitted. As a result she encountered considerable difficulties in trade with her neighbors. However, why not transform this position of weakness into a position of strength? A country which went through underdevelopment, central planning and market organization and which is economically and politically uncommitted might perhaps become a desirable economic meeting place for three different worlds. If so, external convertibility is certainly one of the preconditions for making the mediating role of the Yugoslav market attractive for her partners from the West, the East and the Underdeveloped South (Čičin-Šain, 1968a, p. 82).

V. *Money, Banking and Public Finance*

Banking and Monetary Policy

There has been a lot of experimentation in the Yugoslav economy. This is true for the monetary field more than any other.

Banking can be organized in a centralized or decentralized fashion. Decentralization can be (1) regional, (2) functional or (3) both. Centralization can be (1) absolute or (2) partial. Thus there are five possible organizational solutions. All of them have been tried out at one time or another.

Banking for a Centrally Planned Economy: According to the Institute of Finance, in the socialist economy of 1949 money was a tool used by the state authorities to distribute social product in proportion to the labor of each working-man, to establish economic ties among enterprises and to exercise control over their activities. Money was also a means of accumulation and an instrument of control over plan fulfillment (Finansiski, 1949, p. 63). The banking system was expected to provide money which had such properties.

From pre-war times Yugoslavia inherited a certain number of private and state banks. The former were eliminated by 1947 and the latter were reorganized. The National Bank was a descendent of the Serbian National Bank created in 1883. The former State Mortgage Bank—the heir of a state bank set up in Serbia in 1862 (Uprava fondova)—continued to operate as the State Investment Bank. The Agricultural Bank of 1929 continued to operate in the same field. There was also a Handicraft Bank and, in view of ambitious industralization programs, it appeared advisable to set up a separate Industrial Bank.

The war had not yet been ended when a process of creating regional banks began: six republics—six regional banks.

For a country aiming at central planning, all these banks did not represent a very purposeful arrangement. In September 1946 a consolidation of the banking system began. All existing banks were merged into the National Bank, entrusted with short-term transactions, and the State Investment Bank, which was to deal with investments and foreign loans. Apart from dealing with short-term credit, the National Bank issued currency, performed general banking and agency services for the government and served as a clearing house for the entire economy. In 1948 the two-bank system seemed overly

centralized. Since local enterprises and agricultural co-operatives played special roles at that time, 89 Communal Banks and 6 regional State Banks for Lending to Agricultural Co-operatives were formed. Communal banks were universal banks: they were for servicing local budgets, extending short and long-term credits, collecting savings, controlling plan fulfillments of local enterprises. Banks charged a one percent interest rate which was in fact a commission charge for their services. It was not deemed appropriate for a socialist system to charge interest as a price for capital.

Since it is much easier to control financial transactions conducted via bank accounts than those made in cash, already in 1945 all enterprises and other non-private transactors were obliged to have drawing accounts with the bank. Soon about nine-tenths of payments were conducted without using cash. This was one of the lasting results of the early period of banking development. Payments through bank balances developed into a unique internal payment system, channeled through local offices of the National Bank. It embraced all banks, post offices, enterprises, government funds and a considerable part of the private sector and connected all money streams of the economy into a single consistent system (Vučković, 1963, p. 366).

In many respects the early Yugoslav monetary system was a replica of the Soviet model. This is particularly true for the three instruments of monetary control: credit planning, cash distribution and the automatic collection of invoices.

Credit planning was the only instrument to survive the administrative phase. Until 1950 credit planning simply meant summing up the credit needs of individual enterprises. This was done by planning authorities. The bank was supposed to implement such plans in a routine way. Later, credit plans were transformed into credit balances, which meant that needs were balanced with means. Banks were made responsible for drawing up credit balances (Vučković, 1956, p. 172). The planned amount of credit for individual enterprises was obtained by dividing the output target into an individual capital-turnover coefficient and then subtracting the enterprise's working capital (Vučković, 1963, p. 366).

The main purpose of cash distribution plans was to control recipts (mainly in retail trade, catering and passenger transport) and expenditures (primarily for wages and payments to peasants) made in cash (Stevanović, 1954, pp. 145–46). The cash plan was made for territorial units and for separate money streams and so provided useful information about receipts and expenditures of the population and about various channels in which the money was circulating in the economy. But it was a rather rigid instrument with not much use outside central planning and was therefore abandoned in 1951.

In order to enhance financial discipline, enterprises were forbidden to grant trade credits to each other. The automatic collection of invoices served the same purpose. The bank would automatically credit the seller's account when goods were shipped and then charge the buyer's account. In this way no mutual crediting could be practised. Payments were carried out smoothly. If there was no money in the buyer's account, credit was automatically extended. This, of course, meant that credits would expand beyond the limits set by the credit plan. At first, such matters did not worry planners too much; physical targets and not money flows were important. Other consequences were more disquieting. The total volume of credits depended more on debtors then on banks. The necessary discipline was jeopardized. Sellers did not care about the solvency of their buyers, and also tended

not to pay sufficient attention to delivery terms, assortment and quality of goods. Buyers did not mind accumulating excessive inventories. After a while careless buyers had to be put on "black lists," their drawing accounts were blocked and in many cases they were brought before the courts. The automatic payments mechanism broke down and was in 1951 replaced by free contracts among the trading partners (Vučković, 1957, p. 21).

Learning by Doing: What sort of bank ing system was appropriate for a self-management economy? Centralized or Decentralized? There was a lively discussion on that issue. E. Neuberger surveyed the principal arguments advanced in favor of the one or the other alternative (1959a). Whatever the merits of these arguments might have been per se, the government decided to play safe. No one could be sure of the business behavior of labor-managed enterprises. It seemed advisable that decentralization in the market for goods and services be accompanied by strict centralization in the financial sphere. All other control instruments, remarked J. Pokorn, were to be replaced by bank control and supervision (1956). In March 1952 Communal banks ceased to exist and other banks were merged with the National Bank into one single giant bank with 550 offices and 16,000 employees.

In order to make control as efficient as possible, the working capital of enterprises was transferred to the Bank. Enterprises were to pay a reasonable interest rate, which was to induce them to economize with the credit money.

The shorter-term interest rate was differentiated according to turnover velocity of working capital and ranged from two percent for crude oil production and agriculture to 7 percent for electric power plants. This span was reduced to 5–7 percent in 1953. It was again increased and the rates differentiated in a somewhat different way, for different kinds of credits, in 1954. Experimentation with interest rates continued even later, and in 1956 there were 25 categories of active interest rates (Vučković, 1957, p. 183).

The so-called "social accounting" represented one lasting result of the 1952 reform. The Bank established special accounts—at first thirteen of them—for all important transactions of each enterprise. All changes that took place in the current account of an enterprise were entered here. In this way: the Bank and the government had up-to-date information; the Bank was able to exert stringent control—it would stop any irregular payment, which was particularly important for payments related to wages; the Bank checked the fulfillment of tax and other obligations of an enterprise towards the state. The system was later simplified, the number of separate accounts was gradually reduced and the Bank began to rely more on quarterly accounting statements by the enterprises. A standard accounting scheme, obligatory for all enterprises, made this task a routine matter. In 1959 the social accounting with its drawing accounts system for the entire economy was separated from the National Bank and turned into an independent social service. The work was computerized and the service became very efficient. A little later it was discovered that the Social Accounting Service's monopoly on the payments traffic was not an obstacle to enterprises keeping their financial resources with the banks of their own choice. Today every non-private income earner has a drawing account with the Social Accounting Service, and pays commission charges, and at the same time has a deposit account with one of the banks, and receives interest on deposits.

The proper procedure to be used in extending short-term credits was one of the important problems the all-embracing National Bank had to solve. In those days of romantic beliefs in the possibilities

of inventing simple problem-solving devices—such as the Rate of Accumulation and Funds—that would eliminate the arbitrariness of a bureaucratic apparatus, the Bank hired a couple of mathematicians and asked them to invent appropriate formulae for credit extension. A booklet with several dozens of such formulae was published in 1952 (Miljanić *et al.*, 1956). They were based on turnover coefficients of credits and ratios of sales to costs. Since parameters to be used in formulae could be calculated only as some sort of averages, it was soon discovered that some enterprises got some more credits than they needed, while others badly lacked the money to keep production going. Formulae were abandoned and in 1953 the amount of credit extended was related to the maximum quarterly credit used by the enterprise in the previous year. This favored last-year debtors and penalized good entrepreneurs and had to be abandoned. But the idea of some automatic credit evaluating mechanism was not abandoned.

In 1954 the Bank experimented with credit auctions. Vučković explained that credit auctions were to be a kind of socialist credit market where the supply and the demand of money would meet and determine the general conditions for credit extension (1957, p. 38). The Bank expected that less profitable enterprises would refrain from asking for credit because they would not be able to bear high interest rates. It turned out that precisely the less profitable or unprofitable enterprises were prepared to offer the highest interest rates—up to 17 percent—because they considered credit the only available solution for their problems. The Bank then set the marginal interest rate at 7.5 percent. But this was a negation of the whole idea of auctions. Soon credit was extended automatically to every enterprise that had satisfied the formal conditions of an auction. Since all automatic devices provde inefficient, in 1955 the Bank fell back on the traditional banking practice of an individual evaluation of every credit request.

By 1954 two facts were established: (1) the NES worked well as a whole but (2) the centralized bank left much to be desired. As soon as that had become clear, regional and functional decentralization were initiated. One of the main justifications for decentralization was the socio-economic incongruity between self-management in the commodity market and state monopoly in the financial market. Vučković quoted approvingly the governor of the National Bank, who declared that in a decentralized banking system the credit function would be subject to the control of social self-government instead of bureaucratic management (Vučković, 1957, p. 86). Communal banks with all their diverse activities were re-established. The banks were obliged to keep reserves with the National Bank of up to 30 percent of demand deposits and 100 percent of investment funds. In the next three years three specialized federal banks were added: a foreign trade bank, an investment bank and an agricultural bank. The National Bank was relieved of investment and some other banking operations. Each bank was run by a managing board whose members were partly appointed by the authority that founded the bank and partly elected by the bank's personnel in the proportion 2:1.

After all these changes had taken place, it appeared appropriate to give back working capital to enterprises. This was done in 1956, and the system was stabilized for the time being. Working capital was not given back free of charge; enterprises were obliged to pay an interest rate of six percent.

Banking for a Self-Government Economy: It took eight years before a formerly administratively run economy learned how to handle a few basic financial mech-

anisms. The task of creating an adequate institutional system in the financial sphere was yet to be accomplished. It took eight more years before an outline of such a system became visible.

The deficiencies of the banking system as it developed until 1960 were described by V. Holjevac as follows (1967a). The National Bank offices were inefficient, unimaginative, engaged in distributing the planned increase in credits and executing the decisions of the head office. Communal banks fell under the complete control of local authorities which often made it impossible to conduct a sound business policy of profitable and safe investments. The federal government often directly interfered with the banking business by immobilizing certain kinds of deposits or by running a deficit inconsistent with the social plan. In order to overcome these deficiencies a series of reforms was undertaken. As in the post-1952 period, reforms were carried out in two-year intervals starting with 1961.

In 1961 communal banks became basic and universal credit institutions. In order to eliminate the monopolistic influence of political authorities, a two-thirds majority of members of the banks' managing boards were nominated by workers' councils of the enterprises located in the territory of the bank. Next, eight regional banks reappeared. They were to serve as mediators between communal banks—which were required to keep a 5 percent reserve with respective regional banks—and the National Bank. That was a rather unfortunate arrangement, since it caused the disintegration of the national credit market into six regional markets with different business conditions etc. (Miljanić, 1964, p. 53). This mistake was rectified four years later.

In 1952 an interesting new institution was created. It was called Joint Reserve Funds of the Enterprises. D. Dimitrijević describes Joint Reserves as a semi-financial intermediary. Joint Reserves—created at communal and republican levels—grant credits to those enterprises which have losses, are not competitive, have an unsound financial position and are not eligible for regular bank credit (Dimitrijević, 1968a, p. 19).

For more than a decade Yugoslav banking practice, and monetary theory, maintained a fundamental difference between fixed capital and working capital financing. This made sense in a centrally planned economy, but led to mistaken policies in a market setting. It was now realised that working capital was not homogeneous: it consisted of a constant part, which could and should be financed as fixed capital, and a fluctuating part which was a proper object of short-term credits. In 1961 enterprises consolidated the fixed capital and working capital funds into one single business fund. Thus all liquid assets could be used both for current payments and for capital formation.[8] In order to increase the financial independence of enterprises, they were encouraged to finance the constant part of the working capital out of their own funds and to rely on bank credit for the fluctuating part only. But that was not enough for a full-fledged credit policy.

Policy makers had to solve the following problem: design a flexible credit policy with a minimum of administrative allocations when there is no proper money and capital market. They decided to use so-called qualitative control, which implied regulating the demand for credit. The new

[8] However, enterprises were still obliged to hold five separate accounts, apart from the drawing account, with the Social Accounting Service. These accounts (depreciation, undistributed profits, non-business expenditures, and two types of reserves) were operated under special rules designed to induce enterprises to behave in a proper business fashion (Miljanić, 1966). Separate accounts, of course, reduced the possibility of rational use of money, since it could not be freely transferred from one account to another. However, gradually separate accounts have been eliminated.

policy was introduced in 1963 and one of its architects, N. Miljanić, governor of the National Bank, gave a detailed account of it in a book published a year later (1964). According to Miljanić, final demand ought to be financed out of income produced. This implies that inventory formation should be financed out of accumulation. The Governmental budget deficit could be used as a source of new money, but that is not desirable because in the absence of a money market, the distribution of such money occurs in a haphazard way and cannot be controlled. Miljanić even insisted that the federal budget should be balanced in any case (1964, p. 31). This contention, though clearly not defensible in theory, has some justification in practice in view of the sometimes less then responsible deficit financing of government agencies. The official document of the National Bank adds that in case of a recession it is preferable to increase, selectively, the money supply rather than to run a budgetary deficit (Narodna banka, 1965, p. 28). Since the liquidity trap is non-existent in the Yugoslav economy, this is a valid statement.

New money ought to be used to finance primarily the circulation of commodities. Thus credit is given on the basis of some evidence, invoice or bill of exchange, that a commodity has been sold by a producer or bought by a merchant. Credit cannot be given for sales to final buyers (government, investors, consumers). As exceptions to the rule and on the basis of individual evaluation by the bank, credits can also be given for seasonal stocks and for stocks due to some circumstances beyond the enterprise's control. (In fact, credit for stocks, far from being an exception almost reached the level of credits for commodity circulation) (Miljanić, 1964, p. 72). Apart from this first category of credits, which creates some sort of neutral money, credits can also play an active role in supporting production. Such are credits for specific ventures, primarily for exports, agricultural production and for building apartments for sale. Miljanić also noticed one difficulty with his system. Business operations require that an enterprise always have at its disposal a certain amount of money pure and simple. This money is a part of constant working capital, but, being money, should not be financed out of income. On the other hand, if it is financed by credits, they are clearly not short-term ones. Miljanić feels that revolving credits might do the job (1964, p. 88).

This system lasted for four years and produced some good results. Enterprises knew in advance what conditions they must fulfill in order to obtain credit from the bank. Commercial banks were sure to get credits from the National Bank if they fulfilled the prescribed conditions. But the system was also deficient in many ways. B. Mijović, a director in the National Bank, pointed out that qualitative control (conditions, purpose, duration and kinds of credits) could not quite achieve the aims of quantitative regulation of the money supply. The National Bank had to generate a constant stream of detailed and extensive instructions, which became particularly cumbersome. Since not all practical cases could be envisaged and regulated in advance, the handling of borderline cases caused considerable difficulties. Frequent institutional changes elsewhere in the economy caused additional difficulties (Mijović, 1967, pp. 73, 112). By 1967 the credit system was ripe for a new reform. This time supply of—and not demand for—credit was made a primary object of monetary control. Selective control was accommodated within a system of quantitative regulation.

The three types of credits—investment, commercial and consumer—led to a law providing for the setting up of three types

of banks: investment banks financing fixed and constant working capital, commercial banks extending short-term credit, and savings banks dealing with consumer credit. Table 9 summarizes the latest organizational changes (Basaraba, 1967, p. 78).

Organizational changes reflected very definite policy changes. (1) Federal, Republican and communal banks disappeared. All banks can in principle conduct their transactions over the entire territory of the country. This deterritorialization policy came as a response to frequent complaints against parochialism and unsound political pressures of local and republican authorities. (2) The market orientation of banks resulted in a concentration process that reduced the total number of banks by one half in only three years. By the end of 1968 the number of banks was further reduced to 74. This number ought to be compared with 700 private banks before the war. But the most important was (3), *the change in the setting up and running of the banks.* Here at last, a solution consistent with the organization of the rest of economy was found.

Banks are now established by enterprises and socio-political communities (federal, republican, local) as equal partners. In order to be independent business establishments, banks have their own capital, called the credit fund. The founders invest their capital in the credit fund of the bank and become shareholders. At least 25 founders are required for any bank so as to preserve the essentially service function of the bank. The bank is managed by enterprises and socio-political communities in proportion to the amount of their capital invested in the credit fund. Shareholders are entitled to dividends depending on business success. These dividends cannot be distributed in wages, but can only be used for capital formation. In order to prevent monopolization, no single shareholder can have more than ten percent of the total number of votes in a bank's assembly regardless of the amount of capital invested. Also no enterprise or socio-political community can be refused the right to invest in a bank and take part in its management. The Assembly of a bank consists of investors and representatives of the bank's personnel. It appoints the Executive Committee, the director and his deputy. The Executive Committee implements the bank's general business policy. The Credit Committee deals with individual requests for credit except in some special cases. In order to ensure an objective and expert business evaluation of requests, the Credit Committee is composed of the bank's own experts. The employees of a bank have their own self-management bodies which deal with the distribution of personal income, use of various funds, personnel matters and the like and, through representatives on the Executive Committee and in the Assembly, participate in the management of the bank.

TABLE 9.—BANKS IN YUGOSLAVIA

November 1964		June 1967	
Type of Bank	Number	Type of Bank	Number
Communal banks	206	Commercial banks	61
Republican investment banks	8	Mixed banks	39
Specialized federal banks	3	Investment banks	11
Total	217	Total	111

After a network of commercial banks had been established, the National Bank discontinued its direct business contacts with enterprises and became a central bank in the traditional sense.[9]

In its function of regulating the money

[9] Neuberger examined the role of central banking under three types of economic systems, the Yugoslav system before 1961 being one of them (1958).

supply, in 1961 the National Bank had the following weapons at its disposal (Golijanin, 1967, pp. 95–104).

1. Currency issue.
2. Sales of foreign exchange.
3. Fixing of terms for extension of short-term credit by communal banks.
4. Legally required reserves held by communal (later by commercial) banks with the National Bank. The upper limit was set at 35 percent of liquid deposits.
5. Limits for interest rates (in practice 8–12 percent).
6. Restriction of the use of certain kinds of deposits. This instrument was often and indiscriminately used, which greatly annoyed the owners of funds. I. Perišin points out that in the period 1954–1962 between 34 and 45 percent of total deposits were blocked in this way (1967).
7. Special credits extended by the National Bank to other banks. These credits were used to finance about one half of all short-term credits extended by commercial banks to their clients.
8. Consumer credit policy.
9. Consultations and recommendations.

Compared with traditional banking, some items appear superfluous, but one important item is missing: there is no place for an open market policy since there are, so far, no treasury bills. Instrument 7 is a substitute for that. By special credits new money is created and the liquidity of commercial banks ensured. If a bank wants to reduce excessive liquidity, in order to avoid paying passive interest, it can do so by repaying its credit to the National Bank.

As already mentioned, the 1967 reform replaced credit demand control by credit supply control, and so the functions of the National Bank had to be adjusted accordingly. Instruments 3 and 6 were abandoned and the existing amount of special credits was frozen and could not be increased. Several new instruments were added:

10. Rediscount credit, which is used as an instrument of both global and selective control. It amounts to about 12 percent of all commercial credits. In order to qualify for getting this type of credit a commercial bank must fulfill two conditions: (a) its total indebtedness with the National Bank cannot be greater than its demand deposits; (b) at least one-half of its short-term credits must consist of credits with repayment periods shorter than three months. Condition (b) is a special type of liquidity reserve requirement designed for the Yugoslav environment where there is enormous pressure to use short-term sources for investment loans.
11. Discount rate.
12. Quantitative restriction of credit as an exceptional measure.

This is an impressive array of weapons which, if inappropriately used, can cause considerable damage. In the section on monetary policy we will see how this can happen.

In 1967 a daily market was set up within the Association of Banks as a particular kind of stock exchange for supply and demand of short-term capital. Banks in need can obtain credit for a period not exceeding 15 days (Basaraba, 1967, p. 81). At the time these lines were written, the Federal Parliament passed a package of financial laws providing, among other things, that shares in a bank's capital can be sold to the business public, but not to socio-political communities and to banks themselves. These two events may be con-

sidered as proper beginnings of a stock exchange development—of the Yugoslav variety, of course.

Let me close this section with a note on monetary planning. On the basis of reliable and up-to-date information provided by the Social Accounting Service, a sophisticated system of flow-of-funds accounts was designed. Since 1967 this system has also been used for annual and monthly monetary planning, thus replacing the old-fashioned credit balances. Its author, Dimitrijević, gave a technical description of the methods used in his 1968b article.

Investment Financing: The amount of professional literature on investment financing varies in inverse proportion to the number of complaints against the state of affairs in this field. It is difficult to figure out why this is so. Perhaps it is because investment financing is in a sense a borderline case: neither monetary theorists nor fiscal policy experts nor predominently physical planners feel competent to deal with it. In any case investment financing has been one of the weakest links in economic policy for a long time, and yet no serious study of its problems has been undertaken so far. Thus I will confine the exposition to a description of actual development.

Capital formation may be financed by fiscal means, i.e. out of taxation, or out of enterprises' own funds, by bank loans or by means of securities of various kinds. This is roughly the order in which the various kinds of investment financing have been tried out in Yugoslavia.

Early in 1945 the government created the Fund for Reconstruction whose resources consisted of confiscated war profits[10] and of proceeds from sales of goods supplied by UNRRA. Very soon loans given by the Fund were written off and capital formation was financed in the budgetary fashion typical of a centrally planned economy. Investment resources were allocated by the plan and given to enterprises from the budget free of charge. Enterprises could not sell capital goods; they could only transfer them to other enterprises after having obtained permission to do so. Since the state was the only owner of capital and prices did not matter much anyway, this arrangement was consistent with the rest of the system.

The crucial year of 1952 inaugurated important changes. The Federal budget as a source of investment finance was replaced by the Fund for Basic Capital Development. Investment resources were still allocated without repayment obligations, but the creation of the Fund led to a division of the budget into two separate parts: one was related to administrative expenses and the other consisted of various investment and interventionist funds. This was to become a permanent feature of the Yugoslav budget.

In 1952 the federal government concentrated just about all investment resources in its Fund. That served the purpose of gaining time for the preparation of a more thoroughgoing reform. Already the next year Funds for Crediting Investment Activities were formed. Enterprises established their own investment funds financed out of profits that by the plan were left to them. Both measures led to a considerable decentralization of capital formation financing. The system assumed its more permanent shape in 1954 when Social Investment Funds (SIF) were created at all levels, federal, republic, district and communal. Since then, until the latest reforms, Social Investment Funds were granting loans to business enterprises, while capital formation in the non-business sector (schools, hospitals, government

[10] In a similar setting after the First World War the government had great difficulties in introducing the tax on war profits and, once the required law was promulgated, it could not be implemented (Milojević, 1925, pp. 168–82).

offices, etc.) continued to be financed out of the government budget. The creation of SIF—which tended to multiply as time went by—had an interesting behavioral consequence. Since all levels of the government were under constant heavy pressure to invest, and funds were separated from the budget, their resources tended to be inflated beyond anything envisaged by the Social Plan. In the period 1955–1960 the volume of investment surpassed the target established by the Social Plan by 20 percent (Vasić, 1963, p. 2157).

The reform of 1954 introduced two other important innovations. One consisted of the transfer of capital assets to enterprises. For the privilege of using social capital, they had to pay an interest rate of 6 percent, which was in 1965 lowered to 4 percent. Interest had to be paid on capital used regardless of the source of its finance. The proceeds from this interest as well as the repayments of the loans granted represented resources of the General Investment Fund operated by the federal government. The interest rate on social capital was differentiated according to the aims of price policy and according to the capital intensity of particular industry groups. It ranged from close to zero for agriculture to 1 percent for electric power generation and coal production, to 2 percent for transportation, to 4 percent for ferrous metallurgy and to 6 percent for most other industries. In this way the interest burden, as a percentage of net product, was more evenly distributed among various industry groups. The average rate of interest amounted in 1961 and 1966 to 2.8 and 1.3 percent respectively in terms of capital and to 3.8 and 2.4 percent respectively in terms of net products (Trklja, 1968, p. 23).

The second innovation is related to investment auctions. There are four types of investment allocation decisions: (1) the level of total investment, (2) the allocation of investment funds among sectors of the economy, (3) the allocation among firms within a sector, and (4) the allocation among technological variants within a firm (Neuberger, 1959b, p. 103). The last decision is made by the enterprise, while the first two are determined by the plan. After priorities have been determined, and investment allocated to the various industry groups, the allocation among the enterprises may be carried out by auctions.

This is an old textbook idea. In various texts on socialist economics with neoclassical background one can find statements that run roughly as follows: "In principle, the applicants would be listed according to the level of the rate of interest they offered and if two offered the same rate, the one who offered the shorter period for repayment of the loan would be given preference. The bank would go down the list until the amount allocated for this auction, or category within the auction, was exhausted, and the rate of interest offered by the first intramarginal applicant would become the one that everyone paid." In fact, this is not an invented quotation, but Neuberger's description of actual investment auctions in Yugoslavia (1959b, p. 93). In theory one could, of course, improve this scheme in various ways. One could apply price discrimination in order to siphon out all non-labor income contained in the difference between the offered and paid interest rate, or one could replace point offers by schedule offers. In practice the experiment did not achieve great success. It was soon discovered that the two price criteria—the interest rate and the repayment period—were insufficient. Thus other criteria were added: the percentage share of participation with own resources (differentiated according to industries and ranging from zero for electric power to 80 percent in manufacturing), the shortest period of construction, the lowest cost per unit of output, and regional effects (Vučković, 1963, p. 372; Hanžeković, 1967b, p.

TABLE 10.—THE COMPOSITION OF INVESTMENT IN FIXED CAPITAL BY SOURCE OF FINANCE, EXCLUDING PRIVATE INVESTMENT (IN PERCENTAGES)

	1948	1951	1952	1953	1954	1955	1960	1962	1964	1966	1968
Social Funds and Budgets	99	98	98	87	74	64	52	59	36	16	16
Federation	60	50	95	71	50	47	37	30	7	6	9
Republics	27	41	2	11	12	9	7	9	8	3	3
Communes and Districts	12	7	1	5	12	8	18	20	21	7	4
Work Organizations	1	2	2	13	26	35	37	38	32	46	37
Business	1	2	2	13	26	27	31	30	26	39	31
Non-Business	—	—	—	—	—	8	6	8	6	7	6
Banks	—	—	—	—	—	1	1	3	32	39	47

Sources: For years 1948–1955: *Yugoslav Survey*, 1963, *15*, p. 2167.
For years 1960–1968: *Statistički bilten SDK*, 1969, *3*, pp. 68–69.

220). The main defects of auctions appeared to be the following ones. It takes time and it is very costly to prepare an application for credit. Auctions are held at widely spaced points of time which may not correspond with the enterprise's need for investment funds. As in the case of credit auctions, enterprises were ready to offer high rates of interest just to secure the loan. They did not worry too much about future repayments because the tradition of free social capital was still very much alive and because it looked obvious that a plant of any size cannot be closed down "just because the loan cannot be repaid." Thus the authorities in charge of SIF had to examine every case very thoroughly as they would have had to do even without auctions. According to Neuberger's estimates, at most one-third of all investments at any time were allocated through auctions. In such circumstances auctions gradually degenerated into an old-fashioned administrative distribution of investment from government funds.

Auctions failed. The criteria used for investment allocations from SIF had never been very transparent—another reason for the lack of analytical literature—and had always been greatly influenced by political considerations. As a result "political factories" appeared. All important investment projects were somehow multiplied six times, one for each republic. Besides, Social Investment Funds absorbed two-thirds of total investment resources, and owing to participation requirements, controlled directly an even larger share of total investment. Inefficiency and bureaucratic control were not quite compatible with the self-management aspirations of the economy. Enterprises pressed for an increase in their share in investment finance. The data on actual development in characteristic years are given in Table 10.

A considerable share of investment money in the SIF was obtained through taxation. When in 1962 these "contributions to Social Investment Funds" were raised by 50 percent (Vuksanović, 1966) there was a general outcry against the "expropriation." It was requested that "state capital" be done away with. Two years later the contributions to SIF were abolished, and the funds transferred to bank credit funds. That is why bank investment loans increased so sharply in 1964. The starting principle of the reform of 1965 was: to leave at the disposal of enterprises a larger share of their savings and consequently to restrict the role of socio-political communities in investment decisions (Jovanović, 1965, p. 3222). The pendulum was pushed a little too far in the decentralization direction because it was requested that even large capital in-

tensive projects (power generation, communications) also be financed out of capital concentrated in banks.

The role of the Federation in investment was reduced to the operation of the Fund for Undeveloped Regions that would distribute annually to undeveloped regions close to two percent of national income as investment funds. Republican and communal funds also diminished considerably. But the share of enterprises, with the exception of a short-lived post-reform increase, remained stagnant. As Table 10 shows, what actually happened was that the Federation and the banks simply changed places in investment financing.

In a situation of chronic excess demand for investment resources, banks could easily assume a dominant role. The sum of the regular and penalty rates of interest could be as high as 18 percent. The first recession—which in fact followed the reform—was bound to reduce the investment funds of enterprises and make enterprises more dependent on banks. D. Vojnić points out that in 1968 the repayments of bank loans amounted to 111 percent of net profits of enterprises (1969, p. 89).

With almost one-half of investment resources under their control, banks established themselves as a dominant force in the investment market. What should be done to safeguard the independence of enterprises? The answer is by no means clear. The present discussion has concentrated on possible improvements of the capital market. In 1963 government bonds became negotiable and in 1968 the first enterprise bonds appeared. In 1969 bank shares were invented and the present author has suggested that participating debentures be introduced (1967b). The securities market could supply at least part of the capital outside the bankers' control. Pooling resources and joint ventures are encouraged. After the Social Investment Funds had been abolished, interest on social capital became a mere capital tax that flowed into the government budget. A political decision was taken to abolish this capital tax as soon as possible. It is now being suggested that this interest—amounting to about one-eighth of business investment—be given to enterprises as resources earmarked for investment (i.e., it would be treated similarly to depreciation funds). It will not be surprising if in a little while another reform in this field is carried out. After a money market has been to a certain extent adequately organized, its twin, the capital market, can surely not lag behind for very long.

Anti-Inflationary or (Anti-) Anticyclical Monetary Policy: In a centrally planned economy market disequilibria result in physical shortages; in a market economy they are reflected in inflation. The age-old discussion about the real causes of inflation was resurrected among Yugoslav economists, in particular after 1961.

Monetary theorists, not unexpectedly, tended to see the source of all troubles in an uncontrolled expansion of money supply. M. Ćirović argued that the increased commodity prices represented the way in which the economy adapted itself to an excessive expansion of credit and money supply (1966, p. 183). Similarly M. Vučković believed that inflation was essentially a product of excess demand. Since new money brings along new demand unaccompanied by supply, a market disequilibrium arises and generates increases in prices. The excessive expansion of short-term credits is a consequence of the following deplorable practices: short-term credit is used (irregularly of course) for long-term investments, for non-salable stocks, to cover losses, to finance budget deficits and to finance taxes at all levels of government (Vučković, 1967, pp. 128–29). The last mentioned practice is probably also one of the Yugoslav inventions in the field.

Owing to a fairly completely budgetary decentralization, local governments are very keen on squeezing out of "their" enterprises every possible dinar. In the early days of the NES they could do so by tailoring taxes so as to leave the coffers of the enterprises empty. This phenomenon had been described by Miljanić and Vučković already in 1956 (Miljanić, 1956; Vučković, 1956). Thus in 1954 in one single year, communes managed to increase their budgetary revenues by 98 percent (Vučković, 1956, p. 173). In order to comply with these patriotic requests, enterprises would have to increase prices or ask for credits or both. Credits were readily granted because paying taxes on time had always been considered a first priority. After the budget system had been somewhat more efficiently designed, the arbitrariness in taxation was reduced, but whenever in need communes would simply delay payments for goods and services they bought. In this respect republics and the federation have also been guilty until this very day. It is not surprising, therefore, that the business community does not trust their governments too much and tries to get rid of any "bureaucratic" control.

Now, though it is true that credit was excessive and money supply inconsistent with stable prices, it does not necessarily follow that prices were the consequence and credit the cause in the inflationary process. The hypothesis was tested in the IES and it turned out that there was either no correlation between credit and prices or there was a slight *negative* correlation: higher credits-lower prices. This paradox will become understandable in a moment.

Prices are predominantly determined by changes in wages, and so inflation is most of the time a cost push inflation. As already mentioned in the section on Price Policy, wages appear to be a function of capital intensity, technological rent and institutional monopoly (banks, insurance companies). Wage increases in privileged work-organizations initiate wage increases throughout the economy, and whenever prices cannot bear a cost increase, they are revised upwards. Bajt adds that the high degree of price control increases the pressure of excess demand on the free section of the market, which then generates price increases, and that inefficient investment planning produces an inadequate structure of output which in a semi-closed economy makes it difficult to match demand (Bajt, 1967a; Bogoev, 1967).

Business cycles complicate matters even further. Prices are formed in Yugoslavia in a rather peculiar way. Depreciation and interest on social capital represent fixed elements. Wages, as everywhere, are inflexible downwards. For reasons explained in the next chapter, all taxes are tied up with wages and vary proportionally to wages. Since tax payments enjoy high priority it may happen—and did happen—that the total amount of taxes collected increases in the trough of a depression. Finally, repayments of loans represent an additional fixed element. Thus, as soon as there is a slight retardation of production, the enterprise finds it impossible to cover costs and has to run losses,—or increase prices.

In a downswing a labor-managed enterprise will not dismiss workers. Thus production will be continued and inventories accumulated. Inventory accumulation is financed out of profits and credits. When these two sources are exhausted, involuntary trade credits and price increases will replace them. As far as inflationary pressure is concerned, we may expect price increases in the downswings and stable prices in the upswings. Figure 1 confirms such an expectation.

The analysis just sketched—a result of research of the IES—was unknown at the time monetary reforms were designed and implemented. The traditional view that in-

flation means "too much money chasing too few goods" gained wide acceptance. All one had to do, so it was thought, was to curtail the supply of money and the economy would be stabilized. Stabilization was envisaged exclusively as price stabilization. In the program of the 1965 reform employment targets were not even mentioned. Foreign exchange reform, membership in GATT and co-operation with the IMF were interpreted as an international obligation to keep the dinar stable at all costs—a task which even a Tory government would nonadays be reluctant to undertake, but which was cheerfully attempted in an economy innocently unaware of what it might mean. Tight money policy was to be the only device for achieving price stability. There were some doubts about the wisdom of such a policy, but critics were frowned upon and the policy was implemented. That proved to be fatal. Since prices vary inversely to the cycle, an anti-inflationary monetary policy meant an anti-anticyclical policy, a policy of continuous and direct destabilization.

The vicissitudes of monetary policy in the last eight years have been analyzed by Holjevac (1967b) and Perišin (1969), recently appointed Governor of the National Bank. I will mainly draw on their work and on the research conducted in the IES in the text that follows.

In 1960 the cycle reached the upper turning point (see Fig. 1). That passed unnoticed, but price increases were noticed. The analytical device used in such situations consisted of a comparison of "commodity funds" (social product in real terms) and "purchasing funds" (personal and government consumption and investment in money terms) which good and up-to-date statistics made possible. The differences between the two were interpreted as excessive money supply. In 1960 the difference was considerable and called for monetary restrictions. In addition, in 1961, during the recession, a monetary reform was undertaken with the purpose of instilling business discipline. Enterprises were forced to increase the share of their own funds in total working capital at their disposal. This share was indeed raised from 7.8 percent in 1960 to 22.4 percent in 1961; the operation was financed out of savings which meant less investment (Vuksanović, 1969). Recession was deepened, retail prices continued to rise at a rate of 6-9 percent, inventories accumulated at a rate of 20-25 percent per year, and monetary authorities decided to tighten up the policy. As a result one enterprise after another found it impossible to settle its debts and mutual indebtedness was expanding at a rate of about 50 percent per year. The Federal government ran out of money and obtained a substantial credit which the National Bank; also the "moneylessness" —a new term coined for the occasion—had to be cured by some credit expansion. All this, of course, ruined the credit balances. The year ended with money supply increased at a rate more than twice as high as the one envisaged. Holjevac complained about the absence of monetary discipline and the fact that the National Bank lost control over credit expansion (1967b, p. 36). But as a consequence the cycle was reversed and the rate of growth accelerated.

The upswing continued through 1963 and with all that excessive (from the point of view of monetary planners) money in the economy, prices were remarkably stable; industrial producer prices rose by one, percent retail prices by four percent. By the end of the that year the upswing developed into a boom, industrial output was expanding at a rate of 15-20 percent per year, and the balance of payments deficit was increasing. Several months later the cycle reached the upper turning point and in the second half of 1964 the recession was already in full swing. All symptoms of the 1961 recession were repeated,

and so was the monetary policy. In the second half of 1964 and the beginning of 1965 reserve requirements were raised up to the legal limit of 35 per cent, enterprises were forced to use savings for increases in working capital, investment banks had to use one quarter of their loans for working capital financing, consumer credits were reduced. All this, together with the upheavals caused by the price reform, reduced aggregate demand and output growth from about 15 percent in 1964 to about 4 percent by the end of 1965. Since the tax reform left the Federal government without money, it had to resort to substantial deficit financing, which once again upset monetary planning. But the downswing was arrested for a period of four quarters, and all symptoms normally present at an upwards reversal of the cycle became apparent. However, this time the National Bank had formidable monetary weapons at its disposal and it decided to use them to combat "excessive" liquidity.

For some reason, not explained in the documents, the National Bank established the rule that liquidity reserves of commercial banks held as balances with the National Bank should amount to no more than 6 percent of monetary demand deposits (Perišin, 1969, p. 515). These reserves ran around 10 percent in the second half of 1966. As "purchasing funds" were appreciably higher than "commodity funds"—which is reflected in price increases—it looked obvious that there was too much money in the economy. The National Bank reduced its special credits to commercial banks and put an absolute limit on their credit operations. Consumer credits were further reduced. In 1967 exports were retarded, and so it was decided to depress internal demand even more in order to achieve an export drive. As one might have expected, this did not help exports—in fact their rate of growth was soon reduced below zero in spite of selective export credits—but liquidity reserves were brought down to 5 percent, even lower than planned.

As a result of this anti-inflationary policy output growth was reduced to minus two percent, which had not happened since the Cominform days. The present author estimated losses due to the mistaken monetary policy at eleven percent of social product. Perišin found that gross savings had been reduced from 43 percent of GNP in 1964 to 30 percent in 1967 (1969, p. 517). Unemployment was increasing fast. But price stability was not achieved (see Table 6).

The new system of regulating the money supply proved to be very efficient in reducing money supply to any desired level. This conclusion follows from the foregoing description of its practical operation but can also be illustrated by a series of indices. If we compute ratios of money supply per 100 dinars of transactions, expressed as a sum of gross national product and the output of intermediate goods, we get the following data (Perišin, 1968, p. 63):

TABLE 11

1957	15.3	1963	19.4
1958	15.3	1964	18.3
1959	14.5	1965	14.9
1960	15.1	1966	12.3
1961	15.3	1967	11.3
1962	18.7	1968	13.1

In the three years after 1964 the relative money supply was reduced to 62 percent of its original level. One might be tempted to think that this simply meant increasing the transactions velocity of money. But that is not so; the lack of banking credits was compensated for by involuntary trade credits. The latter amounted to 69 percent of short-term bank credits in 1964, to 138 percent in 1967 and surpassed bank credits almost two times by the middle of 1969.

One other fact is worth noting. Figure 1 shows that boom periods of business cycles

occurred in 1960 and 1963 and recession periods in 1961 and 1967. A glance at Table 11 suffices to see the extent to which monetary policy was cyclically synchronized: there was an abundant money supply in the boom and tight money policy in the recession. Consequently monetary policy has been an important destabilizing factor preventing the economy from exploiting its growing potentials.

The second half of 1967 brought the revival and the acceleration of growth continued through 1968 into 1969. Prices were stabilized for a while, inventories reduced, exports soared in 1969 and monetary policy had a relatively easy job to support these favorable trends. It remains to be seen whether monetary authorities—and monetary theorists—have learned the lesson and whether they will be able to avoid making the same mistakes once the trends are reversed.

Public Finance and Fiscal Policy

Budget for a Centrally Planned Economy: In the first two years after the war the new state tried to make the best of the inherited financial system. Taxation was improved in two ways. Before the war a sales tax levied on consumer goods was a major source of government budgetary revenue. That represented a great burden for poorer sections of the population. Next, income tax progression was mild (up to 32 percent) and there were several separate income taxes for various sources of income. Thus people with several sources of income—i.e. the richer ones—could easily evade paying high taxes. It was only natural that the new revolutionary government would make the necessary corrections. The sales tax was reduced from 62.8 percent of government budgetary revenues in 1939/1940 to 46.5 percent in 1946, and separate income taxes were replaced by a single one applied to the entire personal income at increased rates (Finansiski, 1949, p. 25). However, uneven taxation reappeared soon and even in 1969 B. Jelčić complained that the differential tax burden for the same personal income of different taxpayers meant the negation of principles proclaimed and guaranteed by law (1969, p. 159).

When in 1947 central planning was inaugurated, the financial system of the country had to be changed radically. In the old system the government budget used to finance the work of public administration and some social services. That corresponded to the administrative character of the old state. The new socialist state—as described by the Institute of Finance (Finansiski, 1949, p. 16)—acts as an organizer of the entire economy. The targets are annually elaborated in the economic plan and the budget ought to reflect them financially. Each planning organ has its own budget, which is a constituent part of the overall budget. The sum of all financial plans of all ministries, i.e. of all industries, represents an annex to the budget. Thus the budget becomes the financial plan of the entire economy (Finansiski, 1959, p. 11; Matejić, 1958, p. 170). The budget amounted to 64–83 percent of national income (Perić 1964, p. 126). About one half of budgetary revenues was spent on investments.

R. Radovanović describes four principles on which such a budget was based. (1) Centralization of all resources at the disposal of a political–territorial unit (municipality, district, county, province, republic, federation) in the budget of its government. (2) Financing from the budget of all social activities. As far as business firms are concerned, only net revenues are entered into the budget. (3) Concentration of the budgets of all political–territorial units in the Federal budget to ensure central direction in carrying out the most important tasks. This is the famous principle of budgetary monism. (4) As a result of

(3) funds are allocated among various bodies in accordance with their recognized needs and irrespective of their budgetary potentials. Lower bodies are obliged to implement general policy and higher bodies are expected to provide the necessary resources. This had at least one negative consequence. Lower organs were not stimulated to economize with their funds. Instead of trying to expand production in their territories, they were busy in their budget expenditures and exerting pressure on higher bodies to find necessary resources (Radovanović, 1962, p. 1112).

Taxes in such a system are just a technical means for channeling gross profits into the budget (Tišma, 1964, p. 29). The price of a product consists of cost of production, profit and the turnover tax. Profit is generally a small item and is mostly left to the enterprise. If individual planned profit is higher than the average one, three is extra-profit half of which has to be paid into the budget of the higher administrative organ. A planned loss is covered from the higher budget. If achieved profit is higher than planned, half of the difference is left to the enterprise as an incentive. Turnover tax is just a balancing item in an administratively set price. Since it is charged on all commodities and is paid as soon as a commodity is shipped, it is also used as an indicator of how the implementation of the plan is proceeding.

In order to accommodate productivity change in such a rigid price structure the "decrease in full cost of production" was explicitly planned as a separate item. This decrease is partly paid into the budget and so a rather unusual new type of tax was created. Finally, various types of prices, discussed in the section on administratively set prices generated so-called commercial profit, which was mostly absorbed by the budget.

In 1949 the four items enumerated were (in billions of dinars): turnover tax 66.6, share of profits 4.6, decrease in cost of production 3.8 and share of commercial profits 13.1 (Tišma, 1964, p. 96). Turnover tax represented, of course, the bulk of budgetary revenues.

The major proportion of budget revenues came from the business sector. Taxes paid by the population were steadily decreasing in importance, from 22.4 percent of all revenues in 1946 to 9.7 percent in 1952. As a consequence taxation of the population was governed by extra-fiscal considerations. In 1950 the tax on income earned in the state sector was abolished. (It was to be reintroduced only in 1960). This did not matter much, since wage and salary differentials were greatly reduced and income distribution was extremely egalitarian. But income taxes were retained for the private sector and the progression was rather stiff. For peasants the tax rates went up to 70 percent in 1947 and up to 90 percent in 1948, as compared with the flat rate of 3 percent for the members of peasant work cooperatives (cooperatives organized similarly to state firms) (Finansiski, 1949, p. 34). This tax policy was inspired by the idea of the class struggle and was aimed at inducing peasants to join cooperatives.

The policy of stiff taxation of peasants and artisans was continued also later and for the same reasons. In agriculture it was discontinued after the second agrarian reform in 1953, which reduced the maximum size of agricultural estates to 25 acres and so eliminated any possibility of capitalist development. In 1954 taxation on the basis of cadastral[11] income was introduced, and rates were lowered. Both proved to be stimulating. It is interesting to note that Radovanović described the tax on cadastral revenue as an instrument designed to replace compulsory deliveries while making

[11] Cadastral revenue is the value of the average yield of a specific land category under average weather conditions and using an average land cultivation technique.

sure that a minimum output will be produced (Hanžeković, 1967a, p. 91). There was no possibility of capitalist development in handicrafts either, because artisans could employ at most five workers. However, public opinion held that there was something vicious about private business. Tax rates were substantially reduced only in 1963 (Hanžeković, 1967b, p. 33). The policy of containment continued until the ownership discussions in 1967 analyzed in the section on Ownership Controversy. In the meantime the number of artisan shops was substantially reduced, which caused economic difficulties.

Taxation Experiments: After the French Revolution in 1789,—remarked J. Lovčević, the Constituent Assemly abandoned taxes in favor of contributions. After the Yugoslav Revolution a law on taxes passed in 1946 declared that a tax was "a contribution. . . given to the state for economic development, cultural advancement . . . and for the maintenance of the state apparatus" (Milatović, 1967, p. 34). In spite of all its protests of public finance experts,[12] the term stuck. From 1952 enterprises have been paying contributions (turnover tax representing an exception) and individuals taxes. Contributions somehow emanated from social property, taxes from private property. Since the 1965 tax reform contributions have become synonyms for direct taxes or taxes levied on labor income and the term tax is used to denote various forms of turnover tax or property tax. The terminological confusion did not matter very much. But lack of professional competence in designing an appropriate taxation system did matter. In the period 1952–1965 the tax system was changed five times with obvious consequences as far as the efficiency of conducting business was concerned.

In 1952–1953 the system of AF rates—whose rationale was discussed in the section on Distribution Policy—predetermined the taxation system. Out of accumulation and funds obtained by the application of a rate, prescribed by the social plan, to the net product of an enterprise, the social contribution was paid to the budget. It contained social insurance payments, was proportional to the wage bill and was paid at the flat rate of 45 percent. Wage bills above the standard prescribed were taxed at steeply increasing rates. A tax on extra profits was envisaged by law, but never applied due to technical difficulties (Tišma, 1964, p. 97). Turnover tax was greatly reduced, and amounted to 9–14 percent of budgetary revenues (Jelčić, 1967b, p. 14). Its task was to absorb monopoly profit and to influence price formation (Radovanović, 1953, p. 62).

The system of AF rates helped to eliminate administrative ties between enterprises and planning authorities, but soon degenerated into administrative determination of AF rates for each individual enterprise. It had to be replaced by a system working more in a market fashion. It was not clear how to design such a system. It seemed advisable to make use of the experience of traditional market economies. Instead of net product, profit was the base of taxation for the next four years (1954–1957). Wages became part of costs of production. Profit was taxed at a flat 50 percent rate. The other half of gross profit was used for contributions to SIF's for supplements to basic wages, for enterprise funds and for some other purposes. Wages from profits were linked with contributions to local budgets which amounted to same sort of progressive payroll taxation. A tax on monopoly profit was envisaged but never applied because it proved

[12] Fiscal theory distinguishes taxes, contributions and stamp duties. A tax is a compulsory payment for, in principle, no specific service. A contribution represents a compulsory payment for a specific service and in principle covers the cost. Stamp duty is a payment for a specific service at the initiative of the payer, but it bears no necessary relation to the cost.

impossible to established which part of the income resulted exclusively from the work of the collective. The share of the enterprise (wages and undistributed profits) gradually increased to one third of net product generated (net product included turnover tax) (Tišma, 1964, p. 99).

In this period two interesting new taxes were introduced. Mines, hydroelectric power stations and some other firms were to pay rent. Artisans and peasants were obliged to pay tax on hired labor. The latter tax was insignificant in quantitative terms, because only one-eighth of the artisans and almost no one among peasants hired labor, but served as a reminder that hiring labor meant exploitation.

Wages as part of cost of production were deemed inappropriate for a self-management system. Thus the new system, inaugurated in 1958, was based on the distribution of total enterprise income. That was a switch back from profit to net product, reduced for turnover tax and some other items. There was also a terminological change: wages and salaries were replaced by personal income. With many changes the system lasted until 1964.

The main tax, surpassed only after 1961 by the turnover tax, was the contribution from income. The rates were progressive up to 80 percent. Tax progression was in 1961 replaced by a flat rate of 15 percent and a surtax of 25 percent. In the meantime another development took place. It appeared reasonable to link collective consumption and public services to the level of personal incomes earned in any particular territory. For this purpose contributions to budgets were made out of the wage bill. In 1958 these contributions were progressive, in 1959 a flat rate of 11 percent was charged; the rate was increased to 15 percent in 1963. In 1964 some tax rates were reduced, and mining taxes and contributions to SIF abolished. The abolition of progressive rates led to a reintroduction of the progressive personal income tax (Tišma, 1964, p. 207). This indicated that the economic functions of the payroll tax and personal income tax, as discussed below, had been confused. The share of the enterprise in its net product increased to about one half.

In order to increase this share still further, the last reform of 1965 abolished all contributions from enterprise income. The share in net product jumped to about two-thirds. Since then enterprise taxation has rested exclusively on payroll taxes. If we count social insurance contributions, labor has been made about 60 percent more expensive than necessary. This has serious consequences. Before 1960 taxation created capital saving inducements (Pejovich, 1964): in a labor surplus economy that was rational. Since 1964 taxes have stimulated labor saving practices. Enterprises did in fact react: coal was being replaced by oil, cotton growing and cattle raising by wheat cultivation and so on, and thousands of workers became redundant. Further, flat rates introduced an awkward rigidity and tended to intensify cycles. Finally, the abolition of progressive payroll taxation after 1958 and the lifting of wages control in 1961 meant that two important checks on inflationary pressures were eliminated. We have already discussed the consequences.

Taxation experiments have clearly not been completed. Is there anything one could say about how an appropriate taxation system ought to be designed? On various occasions the IES has made suggestions in this regard, and they may be summed up as follows. The equilization of personal income distribution can be achieved by the familiar progressive personal income tax. There is no need to tax profits, even less to tax them progressively, since capital is socially owned. But there is a need to tax payrolls and to tax them progressively. In order to do this wages

ought to be standardized by applying accounting wages for certain categories of skill. (The skill rating should, of course, not be left to enterprises themselves, just as school diplomas are not issued by pupils themselves.) When faced with the alternative of either losing a greater part of the "excess wage fund" through taxes or using that money for development purposes the working collective will often opt for the latter. This will check wage increases in the most profitable enterprises—which have continually been generating wage pushes—and expand their investment, increasing the supply of their products relative to demand and lowering prices. Labor should be made as cheap as possible (for the enterprise, of course, not for the workers) in order to stimulate labor intensive production. If some taxes still prove necessary, they may be levied on the enterprise income at a flat rate. Such "contributions from income" may be considered as a self-management counterpart to the familiar value-added tax.

While direct taxes have received little attention in professional economic literature, turnover tax has been extensively discussed. And with good reason. It survived through all tax reforms as one of the principal taxes. Since 1954 the share of turnover tax in total budget revenues has oscillated between 29 and 43 percent (Hanžeković, 1967a, p. 28). By 1964 six kinds of turnover tax were in operation (Lazarević, 1965). Producers' turnover tax was inherited from the days of central planning It was levied on some 250 products at rates varying between 2 and 81 percent: it was contained in producers' prices, represented a part of enterprises' gross receipts and was collected at the time the invoice was issued. It was easily and quickly collected, even before bills were paid, and was liked by the government. It was also used as an instrument of price policy. In order to provide independent sources for communal budgets, in 1956 a communal sales tax was introduced. In 1961 owing to the abolition of progression in enterprise income taxation the government ran short of money and introduced the one percent general turnover tax. This was a multiple-stage tax and was intended to reduce the number of middlemen between producers and final consumers: however, apparently no effect of this kind was achieved (Hanžeković, 1967b, p. 47). There was then also purchase tax on specific products, service sales tax and duty on real estate and other transfers.

Producers' turnover tax has been severely criticized. Both the government and the enterprises tended to abuse it as a price formation device. In twelve years its tariff was changed almost one hundred times (Jelčić, 1967c, p. 4). Its handling required a large amount of working capital on the part of the enterprise. It tended to distort prices, and so did the multiple stage general turnover tax. In the case of exports, tax deductions had to be computed and made. For all these reasons the two kinds of turnover tax were abolished and in 1965 replaced by a sales tax levied on consumer goods in retail trade, added to retail prices, charged directly to buyers and collected when the commodity was sold. But a retail trade sales tax cannot be changed often and cannot be differentiated for many products. Thus its use as a price formation instrument is rather limited. It is now primarily a device for collecting budget revenue.

Budget for a Self-Government Economy: A budget is more consistent with a centrally planned economy the more all-embracing it is. Ideally all financial transactions of the economy are to be regulated by the budget. It is the other way round in a self-government economy. Here the budget ought to be restricted to as small a section of the economy as possible in order not to interfere with the economic activi-

ties of work collectives. Ideally the budget should cover only the activities of various state agencies. In this respect the 1952 reform initiated three important developments. They were related to the organization of the non-market sector of the economy, to the creation of various social funds and to the decentralization of budgetary revenues and expenditures.

The Yugoslav tradition had made a sharp division between enterprises (poduzeća) and institutions (ustanove). The former were business establishments, the latter were financed from the budget and roughly corresponded to non-profit institutions in the USA and elsewhere. Since the latter depended on the budget, i.e. on the government administration, for their revenues, it was clear that self-management had little chance of developing. Thus institutions that performed public services and could be financed partly or wholly by selling their services[13] were separated in a special group of "institutions with independent finance." Gradually it became evident that there were two fundamentally different types of public services: the one (government administration, judiciary, police, defense) rendering various administrative services to society, the other (education, science, medical care, etc.) increasing the welfare of the members of society. It seemed appropriate to finance the former from the budget ("public expenditure") and to organize them in a more or less traditional fashion, but the latter ("collective consumption") required a different approach. M. Hanžeković suggested that taxes be used to finance the former and contributions the latter (Hanžeković, 1967a, p. 17).

Next, while there was to be a free market for the short-run operations of enterprises, it appeared advisable to retain substantial central control in the field of capital formation. But capital financing was to be on a credit basis and budgetary financing implied grants without repayments. Thus investment resources were separated from the budget and concentrated in investment loan funds. The budget continued to finance investment projects in the nonmarket sector (schools, hospitals, etc.).

In 1952 social insurance had also been separated from the budget. This decision was motivated by the fact that social insurance could be efficiently operated as an independent social service *under a social self-government regime*. The latter meant that the governing bodies were composed of representatives of various social interests (physicians, social workers, citizens, government representatives).

Very soon there was a proliferation of various funds for housing, for advancement of agriculture and forestry, for roads, for cultural activities, for education etc. Many of these funds had their independent management bodies and obtained their resources from special contributions or from budgets. Hanžeković suggested the following three-fold classification (1967a, p. 13): (1) funds for capital formation (SIF) or for financing public services; (2) funds for financing without repayment obligation or for granting credit; (3) with self-government bodies or without. Definite trends have appeared in further developments. Loanable funds were mostly transferred to the banks. Funds without independent management bodies are used as often temporary instruments of budgetary financing for special purposes. The third category, permanent funds with independent self-government, represents an innovation.

The social-insurance fund set an example. A decade later the example was

[13] The law of 1959 changed this condition into "institutions organized according to the principles of social self-government." The institutions were renamed "independent institutions." In 1965 they obtained the status of work organizations with the same self-management rights as enterprises.

followed by education. At first, T. Konevski remarked, that was just a transmission mechanism in budgetary financing (1968, p. 163). But in 1967 Education Unions were formed to operate the funds. Assemblies at communal and republican levels vote money to be allocated to education funds. Education Unions—self-government bodies composed of representatives of schools, outstanding figures in cultural life, government agencies—distribute the money by negotiating the services to be rendered by various educational establishments. In 1969 Research Unions were formed. They operate funds for research work created in 1960. Unions are shown in the quasi-market sector of Figure 2.

Hanžeković points out that in 1965 funds absorbed 8.8 percent and institutions 14.2 percent of national income, which had to be compared with total budgetary expenditures that amounted to 20.1 percent of national income (1967a, p. 14). Institutions obtain about one third of their income from selling their services to direct buyers (to the market), 50–60 percent of their revenues come from various funds (quasi-market) and only one-tenth derives from budetary subsidies. Such a structure of revenues enabled the non-market (not non-profit, because they do make profits) institutions to gain a considerable amount of independence. Also, they established closer contacts with the buyers of their services and with the rest of the economy. Is there, one might ask, any economic activity in which an Archeology Department of a University, a museum or art gallery can engage? Yes, there is, though perhaps not directly. Tourist agencies and hotels may be, and in fact are, interested in financing the development of an archeological site, a local museum or art gallery. Sometimes these are rather roundabout ways for achieving certain goals, but if they eliminate government control and increase independence, the price may not be too high. Yet there are other costs involved. Konevski points out some of them (1968, pp. 128–65). To administer a fund an administrative apparatus has to be set up. Unlike business enterprises in the market, a school or a hospital is in an inferior position when it negotiates contracts with the funds. Commercialism may and does have detrimental effects in such fields as culture, education, science or medical care. The consumer may be, and often is, victimized. Since it is too early to evaluate the working of the system, one can only invoke the wisdom of the ancient Greeks concerning the organization of human affairs: right proportions, no extremes.

The creation of funds and the establishment of self-financed institutions represent two aspects of decentralization. As a consequence the share of budgetary revenues in national income was reduced from one-third in 1952 to one-fifth in 1967. The third aspect of decentralization was related to the division of revenues among budgets of various socio-political units. The federation was gradually transferring its responsibilities for various social services to republics and communes. As a result the share of federal expenditures in total budgetary expenditures dropped from 74 percent in 1952 to 53 percent in 1968. The trends have been reversed as compared with what happens elsewhere.[14]

The division of budget revenues among various budgets is a somewhat complicated technical problem. Not less than five laws in the period 1952–1965 tried to solve it and with only limited success. In theory there are two possibilities: a separation of revenues, and joint revenues. Both have

[14] In the USA the share of federal revenues in total budgetary revenues increased from 42 percent in 1890 to 75 percent in 1954; in Switzerland federal expenditures amounted to one half of cantonal expenditures in 1913 and to 111 percent in 1958 (Bogoev, 1964, p. 10).

been tried out at one time or another.

After 1952 the budget monism of a centrally planned economy was replaced by a budget pluralism better suited to a self-government economy. The former budgetary system was based on participation in joint revenues, the higher governmental bodies determining the conditions of participation. If lower budgetary units were to be made more independent in the development of revenue sources in their own territories, a system based on a separation of revenue sources seemed more appropriate. Thus sources of revenue were allocated to budgets at various levels. Only the federation was entitled to introduce new taxes, but, if introduced, taxes had to be immediately allocated to specific budgets or funds. In principle every unit was to cover expenditures from its own revenues. This principle was not fully implemented, but there was a great change as compared with the former practice. In two characteristic years republics and communes obtained their revenues in the following ways (Radovanović, 1956a, p. 445):

TABLE 12

	1948	1954
Own revenues	53.7%	72.5%
Participation in joint revenues	43.3%	22.5%
Federal subsidies	3.0%	5.0%

With many changes this system lasted for nine years (1952–1959). Its main shortcomings, as described by Radovanović (1962, p. 115) and K. Bogoev (1964, pp. 188–90) were two. Sources allocated to lower units were not sufficient to meet the recognized needs. Deficits were substantial and were covered by sharing in revenues and by subsidies. These were discussed every year anew, which made lower units very dependent on higher authorities. Next, the lack of objective allocation criteria generated a bargaining process. For both reasons the system failed to provide stability and incentives.

In the period 1960–1964 the budgetary system was again based on participation. Separate sources were allocated only to the federation (they covered 90 percent of its revenues) and to communes about 20 percent of their revenues). Republics and districts had no separate sources. The participation of all units was determined by federal and republican laws. The higher units could not arbitrarily select more favorable sources for themselves. In order to eliminate another source of arbitrariness, participation rates were not differentiated according to sources as before, but instead one single participation rate was applied to all sources of revenue. Participation rates were increased for less developed units, and if this was not sufficient, subsidies were granted. Increased shares and subsidies were to be determined on the basis of the funds needed for carrying out "mandatory tasks and services." However, since objective criteria were not established, the familiar arbitrariness crept into the process. In 1960 only 9 percent, and in the following year only 3.6 percent of all communes were able to cover their needs in the regular way (Bogoev, 1964, p. 205). About one-half of all communes had to rely on both increased participation shares and subsidies. What was intended to be a corrective device turned out to be the main instrument for balancing budgets of lower units.

The 1965 tax reform introduced the separation principle once again. The sources were allocated as follows. Taxes on personal income and sales taxes may be introduced by all socio-political communities. Apart from that, taxes on property (and some other taxes) belonged to communes, estate duties to republics and customs duties to the federation. Communes and republics are empowered to decide independently what kinds of revenue to in-

troduce for their territories and to fix the tax rates. There are two safeguards. The federal government can fix temporarily the limits for the tax rates set by republics and communes. Communes and republics are legally obliged to cooperate with one another in fixing the level of their revenues in order to assure citizens equal treatment. Republics and provinces are entitled to federal subsidies provided their per capita revenue is below the Yugoslav average and they have exhausted all possibilities for collecting revenue through taxation of personal income, in conformity with the economic potential of their population (Turčinović, 1968).

This time the criterion for subsidies has been defined somewhat more precisely. But it has also been criticized. Hanžeković argues that approximately equal budgetary revenue per capita cannot be an appropriate criterion. Instead appropriately defined necessary and justified expenditure should provide a basis for allocations (1967a, p. 7). In fact this seems to be *the* problem of the Yugoslav budget system. Yugoslav territories are extremely unevenly developed. Per capita income in the Republic of Solvenia is 5.4 times higher than in the Autonomous Province of Kosovo. Communal budgetary revenues are, of course, even more unequal: in 1965 the most developed commune in Slovenia obtained per capita revenue almost 16 times higher than the least developed commune in Kosovo. Such extreme differences inevitably ruined all schemes in which allocation criteria were not precisely defined. Konevski complains that in the new system more than one-half of communes in Serbia have to rely on subsidies, which is inconsistent with the philosophy of self-government (1968, p. 116).

In 1968 the government asked a research institute to study the problem. A group under the chairmanship of P. Sicherl prepared a voluminous report (Sicherl *et al.*, 1968). Sicherl finds that although differences between the developed and the underdeveloped regions in per capita income are extreme, differences in nonagricultural income per worker are small. He used a a special statistical method developed by his colleague B. Ivanović (1964) to establish that the distance between developed and underdeveloped regions is appreciably greater in the economic sphere than in the sphere of social services and living standard. In a later article Sicherl argues that it is easier to reduce the distance in the latter sphere (in terms of flows of services) than in per capita national income (1969). As a basis for subsidy computations, Sicherl takes accounting budgetary revenue which he defines as revenue obtained by applying the average Yugoslav tax rates to actual tax sources in the region. The dilemma of whether policy should be based on the equalization of needs or of revenues is resolved in favor of revenues, on the ground that it is difficult to determine needs in an objective way and that to do so is also inconsistent with the philosophy of decentralized decision making. There follows a long and involved discussion of the most appropriate method of determining standard revenue. The difference between the standard and the accounting revenue is to be covered by federal subsidy. Sicherl's Report has been discussed in government and parliamentary committees but has not produced practical results as yet.

Communal Economy: In daily life every man appears in a double capacity: as a producer and as a citizen. Thus direct democracy will also have two aspects: one relating to the work place, the other to the territory where citizens live. As members of working collectives, people engage in self-management. As inhabitants of towns and villages, they manage their affairs by establishing local self-government. The territorial association that corresponds

to the collective at the work place is the commune.

There has been a strong tradition in local government in Yugoslavia since the days of the National Liberation War. People's Liberation Committees, as local government bodies, worked with great independence, initiative and resourcefulness to supply the partisan army and organize daily life in the liberated territories. It is hardly a matter of chance that the first People's Committee and the first Committee of Workers' management appeared simultaneously in the fall of 1941 in the mining town Krupanj. People's Committees continued to exist after the war, but than as components of a rigidly centralized system. The system was based on the principle of democratic centralism, which meant that higher bodies could abrogate decisions of People's Committees.

This practice was radically changed in the fateful year of 1952. The principle of democratic centralism was replaced by the principle of legality control (Dordević, 1957, p. 24). District People's Committees became organs of self-government and Communal People's Committees organs of local government. District Committees had assemblies with two houses: one composed of political representatives, the other of representatives of producers. The next crucial step was taken three years later. The 1955 law on local self-government proclaimed that the Commune was "the basic political-territorial organization of self-government by the working people and the basic socio-economic community of the population on their territory." The Constitution of 1963 changed the phrasing slightly to make the commune "the basic socio-political community." The development of the communal system has been greatly influenced by the historical example set in 1871 by the Paris Commune, "that finally discovered political form in which emancipation of labor can be carried out" (Marx). It is useful to notice, as D. Milivojević points out, that the commune has not been conceived as just a form of otherwise familiar local government. It is a community of those living, working and producing, satisfying their basic needs, and realizing their civil and self-governing rights in a particular territory (1965, p. 8). For a while districts retained certain coordinating functions and then gradually withered away.

Since the commune is a territorial association, one of the first problems to be solved was to determine the size of the territory. The problem was solved by practical experimentation over the period of a decade. Consistent with central planning was a hierarchy of governmental levels. There were three levels below the level of republic: county (oblast), district (kotar) and local committee (mjesni narodni odbor). In 1951 counties disappeared. The orientation towards a market economy made excessive administrative fragmentation—there were more than 7,000 local committees—unnecessary and so in 1952 the number of local committees was halved and committees were replaced by communes. In order to bring local government closer to citizens, in 1955 the commune was made the basic self-government unit. Since, however, the commune was expected to exercise a wide variety of functions, its territory had to be increased. Table 13 depicts the process of territorial transformation. Each new law on territorial changes, remarked E. Pusić, was announced as the last and the definite one (1968, p. 245).

Communal territory was growing larger and larger and by 1967 the average poulation size of the commune (40,000 in 1967) almost reached the poulation size of the district at the beginning of the process (48,000 in 1952). The district became superfluous and disappeared. The larger commune was more efficient, but less self-governing; that is why the new Constitution provided for the creation of local

TABLE 13.—NUMBER OF TERRITORIAL UNITS OF LOCAL GOVERNMENT (END OF THE YEAR)

	1948	1952	1955	1967
Local committees/communities	7967	—	—	4968*
Communes	—	4052	1479	501
Districts	427	351	107	—

* 1965.

Sources: *Jugoslavija 1945–1954*, pp. 35–36. *SGJ–1968*, p. 62. *Yugoslav Survey*, 1965, p. 3296.

communities. These were to be self-governing communities of citizens in rural and urban localities concerned with all activities connected with the satisfaction of the needs of citizens and their families. J. Duričić describes three functions of a local community: it is (a) a form of self-government including traditional political activities, (b) a unit of town planning and (c) an organization taking care of some social services, public utilities, etc. (1965). Pusić is rather skeptical about local communities contributing in any important way to self-government. In his view their activities are too restricted to be particularly attractive to the citizens and in a modern urban setting territorial closeness per se generates no specially active social ties (1968, p. 243). There are 27,706 localities in Yugoslavia, and by 1965 statutes of communes provided for the creation of 4,968 local communities (7.7 percent of communes did not establish local communities at that time). The organizational circle seems to have been closed: communes have replaced districts and local communities have replaced local committees. But considerable social experience has been accumulated in the process.

Apart from exercizing the functions of traditional local government, which include local politics, public utilities, education, social welfare, etc., a commune is also responsible for other aspects of local life. D. Miljković explains this in detail. The commune is expected to harmonize individual and social interests. It is responsible for social property, either under its own control or "belonging" to enterprises. It takes care of economic development and cultural advancement. It coordinates all economic, social and political activities on its territory, prepares a social plan and makes it possible for citizens to participate in the process of social decision-making (Miljković, 1961: Jelčić, 1969). But communal self-government is a contradictory institution, remarked Djordjević, as it carries with it forces of unification and disintegration Both forces will soon make themselves felt.

The 1955 law was preceded by extensive discussions about the functions of the commune. In a paper presented at the annual meeting of Serbian economists in 1954 J. Davičo maintained, and those present agreed, that a labor managed enterprise had no incentive to embark upon substantial capital formation. In his opinion large investment would imply creating a new enterprise which would be equally labor managed and so could not be dominated. For this reason Davičo argued that the commune was "the natural investor in our circumstances" (1954, p. 192). As Table 10 shows, communes indeed became large investors. In 1964, when a maximum was reached, 25 percent of all investment in fixed capital was financed by communes (and districts). Since 1959 communes have been entitled to initiate the setting up of all kinds of enterprises, to bring about mergers or carry out liquidations (Bogoev, 1964, p. 129). However, the last economic

reform put an almost exclusive reliance on enterprises as far as capital formation was concerned, and by 1968 the communal share in investments dwindled to four percent. But this left other economic functions of the commune intact. In cases of failure of an enterprise, the commune shares a good deal of the financial responsibility involved. The commune also gives guarantees for credits and loans granted by the banks to enterprises located on its territory.

For people accustomed to central planning, i.e. to administrative methods in running an economy, it was difficult to imagine a really free market. They were determined to get rid of governmental controls. It seemed obvious that the best way to achieve that was to replace it by communal control. The self-governing commune would tell enterprises what to do and how to behave. In 1954 and 1955 communes were empowered to determine the needs of enterprises and to distribute their profits after federal taxation. Since they were entitled to determine their shares in profits and since they were independent in budget expenditures, communes taxed incomes of enterprises more than the latter could bear. The consequence was a general price rise as shown in Table 6 and Figure 1. In 1956 taxation rights of the communes were again regulated by federal laws (Radovanović, 1956b, pp. 113–16; Bogoev, 1964, p. 166).

Gradually romantic views of conflictless communities, local or otherwise, had to be revised. Hopes have been directed towards an impersonal market mechanism, but expectations have again been a little unwarranted, I am sorry to say as an economist. But at least people were willing to learn from experience. Enterprises gained communal boundaries. Communal banks, which kept appearing in the period 1948–1964, became just commercial banks. The approach to communal economy, self-government and life became far more sophisticated. The actual economic, social and political importance of communes has not decreased, though lately republics show a tendency to encroach upon communal finance.

In an excellent study Bogoev surveys the development of communal finance (1964). In this context one difficult fiscal problem—adequate finance for administrative and, in particular, for social services—may be singled out for closer scrutiny. Bogoev and Petrović point out that the 1957 Resolution of the Federal Assembly on public expenditure and collective consumption which together comprise "general consumption" in Yugoslav terminology as distinct from privately financed consumption) demanded that such expenditure be tied to the economic potentials of the area in question (Bogoev, 1964, p. 179; Petrović, 1968, p. 57). Later the new constitution insisted on the principle of work performed as one of the taxation criteria to be applied to revenues of socio-political units. Tax laws interpreted these two principles to mean that taxes should be collected in proportion to personal income. For this reason the proportional payroll tax gained in importance until after 1964 it became the only tax paid by the enterprises. Since collective consumption is a kind of personal consumption collectively financed, it seemed just and proper to link it with personal incomes earned in a particular territory. The payroll tax was made even more attractive when it was arranged that it be paid into the budget of the commune where people lived and not where they worked or where the enterprise head office was located. It is only recently that the short-comings of the payroll tax and the fallacy in the reasoning by which it was introduced have begun to be discussed.

Let me close this section by a brief re-

view of the main activities of a commune. What communes do is best seen from a breakdown of budgetary expenditures, as shown in Table 14.

Public utilities, education, infrastructural investment and public administration are activities controlled by the commune more than by either republics or federation. Bogoev points out that the communal share in total budgetary expenditures is one of the highest in the world (29–35 percent or 50 percent without defense in Yugoslavia as against 30 percent in Western Germany, 25 percent in Switzerland, 22 percent in Austria and 20 percent or 35 percent without defense in the USA) (1964, p. 329). Whether this share has reached the upper limit remains to be seen.

Fiscal Policy: I add this section for the sake of completeness. But it might as well have been omitted. Strange as it may sound, there is no fiscal policy in Yugoslavia. In fact, this is quite consistent with the belief in the absence—or with the ignorance of the presence—of business cycles.

Fiscal policy can affect aggregate demand via the revenue or the expenditure side of the budget. The revenue side, taxation, has been recognized as a legitimate tool of fiscal policy in theory and is sometimes used in practice. Producers' turnover tax has been occasionally used to affect the general level of prices in order to absorb excessive purchasing power. Otherwise numerous tax changes have been made in order to affect individual prices or to increase the discretionary power of enterprises over their incomes and have not been intended to affect aggregate demand. To a certain extent selective turnover tax reductions have occasionally had price stabilization effects.

The federal government occasionally ran a substantial deficit in recession years, as for instance in 1962 and 1965. But that was purely accidental, a consequence of the combined effects of tax reforms and the lack of revenues. Textbooks on public finance, written invariably by people with training in law, keep on reminding students of the time-honored principle of sound finance: the balanced budget. And since governments on all levels were not too scrupulous in their spending practices, insisting on balancing the budget was quite justified. Bogoev points out that the budget has always been balanced when presented to the Federal Assembly for acceptance and that only in implementation would deficits appear. Deficits have amounted to 10–15 percent of the federal budget and up to 5 percent of republican and communal budgets, but have been much larger for extrabudgetary expenditures (investment, social insurance) (Bogoev, 1966, p. 159).

Table 14.—Budget Expenditure in 1966

	Total expenditure	Federation	Republics and provinces	Communes
Total expenditure effected	100	45.8	19.4	34.8
Education	100	0.1	21.5	78.4
Science and culture	100	5.3	58.1	36.6
Social welfare and medical care	100	52.0	11.6	36.4
Public utilities	100	—	16.2	83.8
Public administration	100	16.7	40.0	43.3
National defense	100	99.7	—	0.3
Infrastructural investment	100	5.3	38.8	55.9

Source: Turčinović, 1968, p. 71.

The first public debate about fiscal policy took place in 1967. At an economic conference in Ljubljana Bogoev (1967), Hanžeković (1967b) and Jelčić discussed the absence of fiscal policy in Yugoslavia and made various suggestions. Bogoev quotes the Resolution of the Federal Assembly on Economic Policy in 1967 which stated that there was excess demand and that not only had all budgets to be balanced but also reserves had to be accumulated. As our Figure 1 shows, Yugoslavia experienced an unusual depression in 1967. Bogoev also points out that proportional tax rates levied on payrolls have cycle-intensifying effects and that the small amount of transfer expenditures (unemployment compensation, debt repayment subsidies) limits the possibilities of an effective anticyclical policy. In the post-war period the federal government raised three internal loans (for the First Five Year Plan, to counteract the effects of the Cominform economic boycott and to finance the rebuilding of Skopje, destroyed by an earthquake). The sole purpose of these loans was to transform a part of personal consumption into investment. Bogoev believes that the rigidity of the existing fiscal pluralism may be softened and an effective anticyclical use made of appropriately designed federal budgetary subsidies to other budgets.

B. Šoškić is the only other economist who has made written contributions related to fiscal policy (1969a). Šoškić was primarily interested in the expansionary effects of public works. In his view the most appropriate objects of increased public financing are: housing and communal construction, road construction, land reclamation and irrigation projects, and power generation projects. Such investment projects are desirable also because of their very low import content, as was pointed out by the IES. Šoškić added that they were also very labor intensive, which is of great importance for a labor surplus economy (1969b).

VI. *Self-Government, Market and Socialism*

Limitations of space preclude discussion of two important lines of economic policy, agricultural policy and regional development policy. But there is one permanent theme of Yugoslav social science discussion which cannot be neglected: the interrelationship between socialism self-government and market. Recent discussions of this problem will be surveyed in this concluding chapter.

I have already discussed the familiar contention that socialism and markets ("commodity production") are incompatible. It was the basis of P. Sweezy's criticism of Yugoslav economic policy as a "gradual transition from socialism to capitalism" (1964). Sweezy argues that the market restricts socialist relations and transforms social ownership into a sort of collective ownership. Material incentives and market orientation necessarily generate a profiteering mentality. The evaluation of social usefulness by profit is characteristic of a capitalist system. Gadgetry and acquisitiveness replace socialist values. This sort of criticism is fairly common. J. Djordjević argues in reply that the undesirable social phenomena are the result of industrial civilization and not only the consequence of the market. The abolition of the market means a return to étatism and state property. Self-government implies free disposal of earned income and, more generally, business autonomy which, in turn, implies markets. If this is not understood, the alternative is an old one: the eschatological idea of state rule and the re-education of man. "Man would be placed under the tutelage of the state (or party, or some other mechanism) to be prepared and educated, so that one day he may become an adult socialist subject" (1966, p. 96).

Yugoslav economists are quite unanimous in believing that the market ought to be maximally exploited as a device of economic organization. Philosophers, however, have their doubts. M. Marković, a leading philosopher actively interested in economic affairs, believes that initial forms of workers' self-management cannot be achieved without material incentives which imply market competition. However, if exclusive reliance on money relations became a permanent feature of the society, self-management might gradually degenerate into a sort of capitalist cooperative. If the results of work were permanently evaluated in terms of income, and if the desire to earn as much money as possible became a permanent and basic interest of a worker, this would produce a personality type not basically different from the type produced by a capitalist society (1965, p. 70).

Referring to Marx, some of my philosopher colleagues declared that socialist commodity production was a *contradictio in adjecto*. In Marx's sense commodity production implies market relationships which result in "commodity fetishism" and various alienation phenomena. I tried to clarify matters in the following way. The familiar statement that commodity production generates capitalism ought to be reversed. Commodity production existed in slavery, feudalism, and capitalism as well as in étatism. It clearly did not determine all these socio-economic systems; on the contrary, it was determined by some more fundamental social relationships and was shaped by respective social systems. Thus, for instance, capitalism resulted from private ownership, étatism from state ownership. Since there are so many types of commodity production, it need not be surprising if we also find socialist commodity production. The elimination of private ownership does not necessarily produce socialism, although it may restrict the role of the market considerably. If private ownership is replaced by state ownership, capitalism is replaced by étatism and commodity fetishism by office fetishism. In both cases relations among people are reified, social inequality preserved, class exploitation continued, essentially human existence made impossible. In socialism social ownership makes social capital equally accessible to anybody while the authoritarianism of a privately managed or a state managed firm is replaced by self-management. In this context the market and planning are not goals but means. If a working collective is to be really autonomous in economic decision-making, the market is indispensable. But planning contradicts the business autonomy of an enterprise and so the choice is between planning and the market—says a time-honored fallacy. In fact social planning, far from restricting, enlarges the autonomy of enterprises for at least three reasons: (1) it reduces uncertainty which is the basic restriction on free decision-making; (2) it increases the rate of growth, the market expands and so the number of available alternatives increases; (3) it equalizes business conditions and so makes the success of a producer less dependent on external conditions which he cannot control and which are economically and socially irrational (Horvat, 1968c).

The nature of the relationship between the market and the plan is a frequently discussed subject. Plan and market have been traditionally contrasted as two separate mechanisms. But some economists try to develop a monistic approach. Bakarić argues that there can be no contrasting, that the law of value reigns supreme and that planning is just one, although the most important, element in it (1963, p. 52). This statement seems to be the reverse of what I said in the preceding paragraph and in the section on decentralization, but the contradiction is more apparent than real. What Bakarić tries to do is to combat the

voluntarism of étatist planning and to show that there is an objectively given framework within which planners are obliged to move. Maksimović understood this statement to mean too much *laissez-faire* to his taste. He criticizes the inconsistencies of the officially proclaimed economic policy and warns that an insufficiently controlled market causes damage to individuals (negation of distribution according to work), and to enterprises (different business conditions in various industries) as well as to the society at large (less than optimal production). All this tends to generate an ideology which maintains that socialism is not economically superior to organized capitalism, that inequality and exploitation are products of human nature and cannot be eliminated (1964).

D. Mišić sees the shortcomings of self-management, as it exists today in Yugoslavia, primarily in the fact that it is confined to the enterprise. Investment resources are not allocated rationally; in the present situation self-management and planning contradict each other, the socialist distribution principle is negated and there is a tendency for group ownership to arise. As a result a *laissez-faire* approach is extolled. Mišić suggests that the self-management structure be completed upwards. He believes that the integration processes, which was discussed in the section on enterprise, are neither fast enough nor quite appropriate. Mišić pleads for an integral system of self-management in which co-ordinating self-management bodies would be created on the level of industries and also regionally. Membership in such associations would be obligatory (1965).

Mišić's system resembles the system of Higher Business Associations which existed in the two-year transitional period 1951-1952. A few years after self-management became operative, the present author suggested a somewhat different approach. A careful study of the economics of the oil industry showed that there was very little to be gained by competition and a lot to be achieved by a co-ordinated policy based on independent and competent research. I suggested that industries possesing similar characteristics establish common but independent economic-technological research institutes. The institutes would prepare alternatives for major policy decisions. The most acceptible alternative, perhaps modified in the process, would be chosen by the representatives of enterprises through some sort of self-management mechanism. The industrial research institutes would also serve as development planning institutions and as such would co-operate with territorial planning bureaus (Horvat, 1962c, ch. 24).

Self-management in enterprises is just one element in an integral system of social self-government. Pusić points out that such a system has three basic components: territorial (various levels of government); functional (enterprises and institutions, i.e., work organizations); and social (cultural, religious and other associations of individuals). Pusić is mostly concerned with the first component. He is thus the first among Yugoslav authors to study systematically the problem of the withering away of the state—generally considered utopian outside Yugoslavia. The state will wither away when government over individuals is replaced by the management of things. Engels took this famous phrase over from Saint-Simon. The latter, as well as other writers of his time, maintained that public administration was exclusively an instrument of power but that it was otherwise unimportant for the life of a nation. Marx and Engles argued with the first part of the statement, but regarded public administration as very important. Later an important duality appeared: public administration was no longer exclusively an instrument of power, but was also en-

trusted with various socially necessary activities: education, medical care, social welfare etc., basically differ from defense, police and judiciary. The monopoly of physical power might occasionally be useful is not at all necessary when social services are concerned. In socialism public administration without state political power becomes the question of the day. In other words, systematic planning and coordination of social services does not presuppose any longer the existence of a commanding center such as is political power (Pusić, 1968). The interest unions and the quasi-market, discussed in the section on institutional framework, represent an attempt to move in this direction.

Self-government is not a purely economic phenomenon. While economists are, naturally enough, primarily interested in economic aspects, other social scientists explore additional dimensions. Lj. Tadić, the political scientist, points out that Yugoslav self-government socialism is mostly confined to the economic sphere. It has been developed on the micro level without a corresponding reflection on the macro level, that of the global society (Simpozij, 1969, p. 55). S. Stojanović, the philosopher, maintains that without faster political democratization it is impossible to create self-government on higher levels of social organization (Simpozij, 1969, p. 34). R. Supek, the sociologist, explains that political pluralism does not mean a multi-party system which can also be bureaucraticized. In a self-government setting political pluralism means direct control of various centers of power. How this is to be achieved is an open problem. Supek expects a certain duality of power to develop at first, a combination of classical representative democracy and self-government.

Evidently, self-government is not a closed and complete system. Many questions are still open, many problems unresolved. The Yugoslav social laboratory is bound to be active for some time to come.

References

D. Anakiovski, "Foreign Trade in the Yugoslav Reform," *Yugoslav Survey*, 1969, *3*, 71–84.

D. Avramović, "Funkcija deviznog kursa u socijalističkoj privredi," *Ekonomist*, 1952, *3*, 3–31.

A. Bajt, "Osebni donodki in delovna storilnost," *Ekonomska revija*, 1956, vol. VII, 97–134.

———, *Raspodela nacionalnog dohotka i sistem ličnih dohodaka u našoj privredi*, Beograd 1962.

———, "Optimalna veličina investicija iz nacionalnog dohotka," *Ekonomist*, 1958, vol. XI, 79–91.

———, "Stopa rasta u nacrtu perspektivnog plana," *Ekonomist*, 1963, vol. XVI, 584–91.

———, "Izvori inflacije u razdoblju posle reforme," *Ekonomist*, 1967a, vol. XX, 141–46.

———, "Faktori dohotka i osnovne ekonomske zakonitosti u njegovoj raspodjeli u socijalističkoj tržišnoj privredi," *Ekonomist*, 1967b, vol. XX, 347–87.

———, "Yugoslav Economic Reforms, Monetary and Production Mechanism," *Economics of Planning*, 1967c, vol. VII, 201–18.

———, "Društvena svojina–koektivna i individualna," *Gledišta*, 1968, vol. XIX, 531–44.

———, "Fluctuations in Growth Rates in Post War Socialist Economies," *International Economic Seminar—CESES*, Balatonfüred 1969a.

———, "Privredna kretanja i ekonomska politika u 1969. i 1970. godini," in *Aktuelni problemi ekonomske politike Jugoslavije, 1969/1970*, Zagreb 1969b, pp. 5–17.

V. Bakarić, *Problemi zemljišne rente u prelaznoj etapi*, Zagreb 1950.

———, *Aktuelni problemi izgradnje našeg privrednog sistema*, Zagreb 1963.

———, *Aktuelni problemi sadašnje etape revolucije*, Zagreb 1967.

P. Basaraba, "Changes in the Organization and Management of Banks," *Yugoslav survey*, 1967, *4*, 77–81.

V. Begović, Dvije i po godine Petogodišnjeg plana," *Komunist*, 1949, *5*, 82–101.
E. Berković, "Differentiation of Personal Incomes," *Yugoslav Survey*, 1969, *1*, 81–90.
R. Bićánić, "Economic Growth Under Centralized and Decentralized Planning: Yugoslavia—A Case Study," *Economic Development and Cultural Change*, 1957, vol. V, 63–74.
———, *Ekonomska politika Jugoslavije*, Zagreb 1962a.
———, "The Threshold of Economic Growth," *Kyklos*, 1962b, vol. XV, 7–28.
———, "Centralističko, decentralističko ili policentričko planiranje," *Ekonomist*, 1963a, vol. XVI, 456–69.
———, "O monocentričnom i policentričnom planiranju," *Ekonomski pregled*, 1963b, vol. XIV, 469–528.
———, "Economics of Socialism in a Developed Country," *Foreign Affairs*, 1966, vol. 44, 633–50.
R. Bičanić, *Problems of Planning: East and West*, The Hague 1967.
D. Bilandžić, *Management of Yugoslav Economy: 1945–1966*, Beograd 1967.
———, "Odnosi izmedje samoupravljanja i rukovodjenja u poduzeću," in *Savremeno rukovodjenje i samoupravljanje*, Beograd 1969, pp. 67–96.
D. Bjelogrlić, "O nekim problemima društvenog usmeravanja privrede," in *Usmeravanje društvenog razvoja u socijalizmu*, Beograd 1965.
K. Bogoev, *Lokalne finansije*, Beograd 1964.
———, "Opšti prikaz fiskalnog sistema i fiskalne politika Jugoslavije," *Univerzitet danas*, 1966, *9–10*, 149–63.
———, "Stabilizaciona fiskalna politika," Ekonomist, 1967, vol. XX, 1–28.
D. Čalić, *Metodologija planiranja proizvodnje*, Beograd 1948.
D. Čehovin, *Ekonomski odnosi Jugoslavije s inostranstrom*, Beograd 1960.
F. Černe, *Planiranje in tržni mehanizem v ekonomski teoriji socijalizma*, Ljubljana 1960.
———, *Tržište i cijene*, Zagreb 1966.
———, "Poskus ekonomsko–logičnega testiranja sedem hipotez iz teorije dohodka," *Ekonomska revija*, 1967a, vol. XVIII, 12–29.
———, "O stabilizaciji in nihanjih v gospodarstvu," *Ekonomska revija*, 1967b, vol. XVIII, 212–29.
A. Cičin-Sain, *Devizni režim i konvertibilnost dinara*, Zagreb 1967.
———, "Problemi konvertibilnisti dinara," *Ekonomist*, 1968a, vol. XIX, 79–102.
———, "Fiksni ili fleksibilni kursovi," *Ekonomist*, 1968b, vol. XIX, 642–48.
M. Ćirović, *Novac i kredit*, Beograd 1966.
N. Čobeljić, *Politika i metodi privrednog razvoja Jugoslavije*, Beograd 1959a.
———, "Tri osnovna problema u teoriji razvoja nedovoljno razvijenih zemalja," *Ekonomist*, 1959b, vol. XII, 225–53.
N. Čobeljić, K. Mihajlović and S. Djurovic, "Problem našeg tržišta s naročitim osvrtom na tržište poljoprivrednih proizvoda," *Ekonomist*, 1954, *3–4*, 31–70.
——— and R. Stojanović, *Teorija investicionih ciklusa u socijalističkoj privredi*, Beograd 1966.
G. D. H. Cole, *Guild Socialism Re–Stated*, London 1920.
S. Dabčević *et al.*, *O nekim problemima privrednog sistema*, Zagreb 1962. (Reprinted in *Ekonomski pregled*, 1963, *3–5*.)
S. Dabčević–Kučar, "Decentralized Socialist Planning: Yugoslavia," in E. E. Hagen, ed., *Planning Economic Development*, Homewood, Ill. 1963. pp. 183–222.
M. Dautović, "Economic Integration," *Yugoslav Survey*, 1968, *2*, 75–82.
J. Davičo, "Privredni problemi komune," *Ekonomist*, 1954, *3–4*, 185–95, discussion, 195–208.
D. Dimitrijević, "The Financial Structure in a Changing Economy: the Case of Yugoslavia," *Florida State University Slavic Papers*, 1968a, vol. II, 1–22.
———, "The Use of Flow–of–Funds Accounts in Monetary Planning in Yugoslavia," *Review of Income and Wealth*, 1968b, Series 14, 101–116.
J. Djordjević, "Teorijska i ustavna pitanja planiranja u Jugoslaviji," in *Usmeravanje*

društvenog razvoja u socijalizmu, Beograd 1965. pp. 7–28.

———, "A Contribution to the Theory of Social Property," *Socialist Thought and Practice, 1966*, 24, 73–110.

M. Dobrinčić *et al.*, *Privredni sistem FNRJ*, Zagreb 1951.

A. Domandžić, "Customs Tariff," *Yugoslav Survey*, 1966, 24, 3485–88.

E. Domar, "The Soviet Collective Farm," *American Economic Review*, 1966, vol. LVI, 734–57.

J. Dordević, *Sistem lokalne samouprave u Jugoslaviji*, Beograd 1957.

———, The Communal System in Yugoslavia," *Annals of Collective Economy*, 1959, vol. XXX, 169–207.

———, "A Contribution to the Theory of Social Property," *Socialist Thought and Practice*, 1966, *24*, 73–110.

A. Dragičević, *Potreban rad i višak rada*, Zagreb 1957.

———, S. Štampar and B. Horvat, *Naše Teme*, 1962, vol. VI, 872–94, 1318–33, 1487–1523; 1963, vol. VII, 99–100.

I. Drutter, "Uticaj koncentracije ponude na cijene i poslovni uspjeh privrednih organizacija," *Ekonomist*, 1964, vol. XVII, 697–700.

———, "Tržišni aspekti koncentracije," in Ekonomski institut, *Ekonomske studije 3*, Zagreb 1965.

———, "Sistem cijena i tržišnih odnosa," in ed. *Poduzeće u reformi*, Zagreb 1968. pp. 95–132.

D. Dubravčić, *Ponašanje samoupravnog Poduzeće u reformi*, Zagreb 1968. pp. 95–132.

———, "Prilog zasnivanju teorije jugoslavenskog poduzeća: Mogućnosti uopćavanja modela," *Ekonomska analiza*, 1968, vol. II, 120–27.

U. Dujšin, "Determinante izbora izmedu fiksnog i fleksibilnog kursa kod nas," *Ekonomist*, 1968, vol. XXI, 592–98.

J. Duričič, "Local Communities," *Yugoslav Survey*, 1965, vol. VI, 3287–300.

K. Džeba and M. Beslać, *Privredna reforma*, Zagreb 1965.

I. Fabinc, "Uloga carinske politike u zemljama u razvoju," *Medjunarodni problemi*, 1963, *4*, 27–39.

———, et al., "Problemi ekonomskih odnosa s inozemstvom," in ed. *Poduzeće u reformi*, Zagreb 1968a, pp. 133–216.

———, "Elementi programa zaštite jugoslavenske privrede," *Ekonomist*, 1968b, vol. XXI, 41–60.

A. Fiamengo, "Samoupravljanje i socijalizam," in Janičijević, ed., *Društveno samoupravljanje u Jugoslaviji*, Beograd, 1965, pp. 11–38.

W. Friedmann and L. Mates, eds., *Joint Business Ventures of Yugoslav Enterprises and Foreign Firms*, Belgrade 1968.

M. Frković, "Disparitet spoljnotrgovinskih kurseva u našoj privredi," *Ekonomist*, 1957, vol. X, 79–97.

A. Gams, "Društvena svojina i društveno usmeravanje," in *Usmeravanje društvenog razvoja u socijalizmu*, Beograd 1965, pp. 50–67.

M. Golijanin, "Credit and Money Control," *Yugoslav Survey*, 1967, *3*, 93–104.

D. Gorupić, and I. Perišin, "Proširena reprodukcija i njeno financiranje," *Ekonomski pregled*, 1965, pp. 109–30.

D. Gorupić, "Tendencije u razvoju radničkog samoupravljanja u Jugoslaviji," *Ekonomist*, 1967, vol. XX, 593–638.

———, "Samoupravno poduzeće i privredna reforma," in *Poduzeće u reformi*, Zagreb, 1968, pp. 3–26.

———, "Razvoj samoupravnih društvenih odnosa i samoupravno odlučivanje u privredi," *Ekonomiski pregled*, 1969, vol. XX, 1–26.

V. Guzina, "Medunarodni zajmovi i socijalistička izgradnja," *Komunist*, 1950, *6*, 21–79.

M. Hanžeković, *Problemi društvenih financija*. Mimeograph, Institute of Economics, Zagreb 1967a.

———, "Djelovanje porezne i monetarnokreditne politike na stabilizaciju jugoslavenske privrede," *Ekonomist*, 1967b, vol. XX, 29–49.

V. Holjevac, *Kreditno-monetarni problemi*. Mimeograph, Institute of Economics, Zagreb 1967a.

———, *Kreditno monetarni problemi 1960–1967*. Mimeograph, Institute of Economics, Zagreb 1967b.

B. Horvat, "Još jedan prilog pitanju prelaznog perioda," *Ekonomist*, 1951, *5–6*, 45–56.

———, "O problemu rudničke rente," *Ekonomski pregled*, 1953, vol. IV, 253–57.

———, "The Optimum Rate of Investment," *Economic Journal*, 1958, vol. LXVIII, 747–67.

———, and V. Rašković, "Workers' Management in Yugoslavia: A Comment," *Journal of Political Economy*, 1959, vol. LXVII, 194–98. B. Ward, "Reply," 199–200.

———, "A Restatement of a Simple Planning Model with Some Examples from Yugoslav Economy," *Sankhya*, Series B, 1960, 29–48.

———, *Towards a Theory of Planned Economy*, Beograd 1964. (Serbo-Croatian ed., 1961.)

———, ed., *Uzroci i karakteristike privrednih kretanja u 1961. i 1962. godini*, Beograd 1962a.

———, "Raspodela prema radu medu kolektivima," *Našta stvarnost*, 1962b, vol. XVI, 52–66.

———, *Ekonomika jugoslavenske naftne privrede*, Beograd 1962c.

———, *Note on the Rate of Growth of the Yugoslav Economy*, Beograd 1963.

———, *Samoupravlenie, centralizm i planovanie*, Beograd 1964.

———, "The Optimum Rate of Investment Reconsidered," *Economic Journal*, 1965, vol. LXXV, 572–76.

———, "Prilog zasnivanju teorije jugoslavenskog poduzeća," *Ekonomska analiza*, 1967a, vol. I, 7–28.

———, "Jugoslavenski sistem samoupravljanja in uvoz tujega kapitala," *Ekonomska revija*, 1967b, vol. XVIII, 406–17.

———, *Ekonomska nauka i narodna privreda*, Zagreb 1968a.

———, "An Integrated System of Social Accounts for an Economy of the Yugoslav Type," *Review of Income and Wealth*, 1968b, Series 14, 19–36.

———, "Socijalistička robna proizvodnja," *Gledišta*, 1968c, vol. XIX, 1321–30.

———, *Ogled o jugoslavenskom duršitvu*, Zagreb 1969a. Eng. ed., *An Essay on Yugoslav Society*, New York 1969a.

———, "Technički progres u Jugoslaviji," *Ekonomska analiza*, 1969b, vol. III, 29–57.

———, "Planning in Yugoslavia," paper presented at the conference on Crisis in Planning, University of Sussex, 1969c.

———, *Primjena medjusektorske analize u planskom bilanciranju privrede*, Beograd, 1969d.

———, et al., *Integrirani sistem društvenog računovodstva za jugoslovensku privredu*, Beograd 1969e.

———, *Privredni ciklusi u Jugoslaviji*, Beograd 1969. Eng. ed., *Business Cycles in Yugoslavia*, New York 1970

B. Ivanović, *Primena metoda I-odstupanja u problemima odredjivanja ekonomske razvijenosti*, Institut ekonomskih nauka, separat br. 13, Beograd 1964.

M. Janković, "Lični dohoci kao faktor podizanja životnog standarda," in *Obračun i raspodela osobnih dohodaka u radnim organizacijama*, Zagreb 1968, pp. 155–68.

B. Jelčić, "Poreski instrumenti kao instrument ekonomske politike," *Ekonomist*, 1967a, vol. 50–63.

———, *Problemi društvenih financija/prihoda/*. Mimeograph, Institute of Economics, Zagreb 1967b.

———, "*Ekonomski učinci oporezivanja prometa proizvoda*. Mimeograph, Institute of Economics, Zagreb 1967c.

———, "Poreska i budžetska politika," in *Aktuelni problemi ekonomske politike Jugoslavije 1969/1970*, Zagreb 1969.

B. Jelić, "Neki aspekti dejstva plana i tržišta u našoj privredi," *Ekonomist*, 1958, vol. XI, 183–201.

———, "Characteristics of the Yugoslav Economic Planning System," *Socialist Thought and Practice*, 1961, *1*, 59–81.

———, *Sistem planiranja u jugoslavenskoj privredi*, Beograd 1962.

B. Jovanović, "Reform of the Credit and Banking System," *Yugoslav Survey*, 1965, *22*, 3216–236.

P. Jurković, "Suština i značaj promjena u sistemu utvrdjivanja i raspodjele dohotka," *Ekonomski pregled*, 1969, vol. XX, 27–52.

E. Kradelj, *et al.*, *Razvoj privrede FNRJ*, Beograd 1956.

———, "Basic Principles of the New Constitution," *Yugoslav Survey*, 1962, *11*, 1529–56.

B. Kidrič, *Privredni problemi FNRJ*, Beograd, 1948.

———, "Kvalitet robno–novčanih odnosa u FNRJ," *Komunist*, 1949, *1*, 33–51.

———, *Privredni problemi FNRJ*, Beograd 1950a.

———, "Teze o ekonomici prelaznog u našoj zemlji," *Komunist*, 1950b, *6*, 1–20.

———, "O nekim teoretskim pitanjima privrednog sistema," *Komunist*, 1952, 41–67.

———, *Sabrana dela*, Knjiga III, Beograd 1960.

T. Konevski, *Fundamentalnost i razvojne smernice novog sistema finansiranja društveno-političkih zajednica*, Beograd 1968.

M. Korač, "Prilog pitanju o prelaznom periodu," *Ekonomist*, 1951, *3–4*, 37–46.

———, *Analiza ekonomskog položaja privrednih grupacija na bazi zakona vripednosti*, Zagreb 1968.

M. Košir, The *Kranj Commune*, Beograd 1966.

O. Kovač, *Uzroci i posljedice struckturne neravnoteže u platnom bilansu Jugoslavije*. Mimeograph, Institute of Economic Studies, Beograd 1966.

P. Kovač and Dj. Milijević, *Samoupravljanje proizvodjača u privredi*, Beograd 1958.

M. Kovačević, "Enterprise Rules and Regulations," *Yugoslav Survey*, 1969, *1*, 1–8.

S. Kraigher, "O politični ekonomiji v prehodnom razdoblju," *Ekonomska revija*, 1950, *1–2*, 9–46.

I. Lavrač, "Konkurencija i stimulacija u našem privrednom sistemu," *Ekonomist*, 1958, vol. XI, 601–19.

———, "Cena upotrebne vrednosti kapitala," *Ekonomska misao*, 1968, vol. I, 407–23.

B. Lazarević, "Turnover Tax," *Yugoslav Survey*, 1965, vol. VI, 3311–20.

G. Lêman, *Stellung und aufgaben der ökonomischen Unternehmungen*, Berlin 1967.

———, *Ungelöste Fragen in jugoslawischen System der Arbeiterselbstverwaltung*, Köln 1969.

F. Lipovec, "Razvoj profitne mere v sistemu samouprave delovnih kolektivov," *Ekonomska revija*, 1954, vol. V, 141–51.

S. Lovrenović, *Ekonomska politika Jugoslavije, Sarajevo* 1963.

G. Macesich, *Yugoslavia. The Theory and Practice of Development Planning*, Charlottesville, Va. 1964.

Lj. Madžar, "Jedna empirijska analiza stabilnosti spoljnotrgovinskih tokova," *Ekonomist*, 1968, vol. XXI, 580–87.

V. Majhsner, "Intervalutarni kurs i cene," *Ekonomski anali*, 1956, *3*, 186–204.

V. Majksner, "Intervalutarni kurs i cene,"

I. Maksimović, *Teorije socijalizma u gradjanskoj ekonomskoj nauci*, Beograd 1958.

———, "Razmišljanja o nekim teoretskim i idejnim pitanjima robne proizvodnje povodom našeg privrednog sistema," *Ekonomist*, 1964, vol. XVII, 209–26.

E. Mandel, "Yugoslav Economic Theory," *Monthly Review*, 1967, *11*, 40–49.

M. Marković, "Socijalizam i samoupravljanje," in *Smisao i perspektive socijalizma*, Zagreb 1965, pp. 54–71.

T. A. Marschak, "Centralized versus Decentralized Resource Allocation: The Yugoslav Laboratory," *Quarterly Journal of Economics*, 1968, vol. LXXXII, 561–87.

K. Marx, *Rani radovi*, Zagreb 1953.

M. Matejić, *Javne finansije*, Beograd 1958.

V. Medenica, "A Survey of the Major Results Achieved in the Implementation of Yugoslavia's 1966–1970 Social Plan," *Yugoslav Survey*, 1968, *4*, 27–46.

M. Mesarić "Prilog diskusiji o obliku gravitacione cijene u socijalistickoj, privredi," *Ekonomski pregled*, 1965, vol. XVI, 607–34.

———, *Planiranje privrednog razvoja*, Zagreb 1967.

———, "Uloga planiranja u jugoslavenskom privrednom modelu," *Ekonomist*, 1969, vol. XXII, 403–26.

P. Mihajlović and S. Tanović "Veza jugoslovenskog izvoza s konjukturom u svetu," *Ekonomist*, 1959, vol. XX, 45–79.

B. Mijović, *Novčana i kreditna politika*, Beograd 1967.

S. M. Milatović, *Poreski sistem*, Beograd 1967.

V. Milenković, "Spoljna trgovina," in *Razvoj privrede FNRJ*, Beograd 1956, pp. 399–399–419.

M. Miletić, "Da li je upravni odbor prevazidjen," *Direktor*, 1969, *9*, 56–60.

R. Milić, *Ekonomika FNRJ*, Beograd 1951.

N. Miljanić, *et al.*, *Kreditni i finansijski sistem u Jugoslaviji*, Beograd 1956.

———, "Pilog izučavanju problematike novca," *Ekonomski pregled*, 1956, vol. VII, 12–24.

———, *Novac i kredit*, Zagreb 1964.

———, "Reguliranje monetarnog volumena u SFR Jugoslaviji," *Univerzitet danas*, 1966, 9–10.

D. Milivojević, *The Yugoslav Commune*, Beograd 1965.

Dj. Miljević, *Privredni sistem Jugoslavije*, Beograd 1965.

D. Miljković, "Komuna i društvena reprodukcija," in *Privredni sistem i ekonomska politika Jugoslavije*, Beograd 1961, p. 66.

D. M. Milojević, *Neposredni porezi Srbije i Kraljevine Srba, Hrvata i Slovenaca*, Beograd 1925.

D. Mišić, "Sistem integralnog samoupravljanja u jugoslavenskoj privredi," *Ekonomist*, 1965, 289–312.

P. Mitić, "Ekonomske integracije, svjetsko tržište i Jugoslavija," *Gledišta*, 1969, vol. XX, 1073–86.

Z. Mrkušić, *Medunarodna trgovina i trgovinska politika*, Beograd 1963.

———, "Neka pitanja na alternativu: prilagodjavavje deviznog kursa—direktna kontrola," *Ekonomist*, 1967, vol. XX, 89–102.

———, "Problemi prilagojavanja deviznog kursa," *Ekonomska misao*, 1969, vol. II, 133–41.

———, *Spoljnoekonomska politika Trećeg svijeta*, in press.

E. Neuberger, "The Role of Central Banking under Various Economic Systems," in C. J. Friedrich and S. E. Harris, eds., *Public Policy*, Cambridge 1958, pp. 227–54.

———, "Centralization vs. Decentralization: The Case of Yugoslav Banking," *American Slavic and East European Review*, 1959a, vol. XVIII, 361–73.

———, "The Yugoslav Investment Auctions," *Quarterly Journal of Economics*, 1959b, 88–115.

D. Nikolić, ed., *Elementi metodologije planiranja dugoročnog privrednog razvoja*, Beograd 1964.

M. Novak, "O prelaznom periodu," *Ekonomski pregled*, 1952, vol. III, 203–13.

———, *Uvod u političku ekonomiju socijalizma*, Zagreb 1955.

———, *Organizacija poduzeća u socijalizmu*, Zagreb 1967.

S. Obradović, *Uvod u analizu spoljne trgovine*, Beograd 1962.

A. Papić, "Investment Financing in Yugoslavia," *Annals of Collective Economy* 1959, vol. XXX, 208–31.

N. Pašić, *Javne korporacije u Velikoj Britanij i drugim zapadnim zemljama*, Beograd 1957

M. Pečujlić, *Klase i savremeno društvo*, Beograd 1967.

S. Pejovich, "Taxes and Pattern of Economi Growth: the Case of Yugoslavia," *Cahier de l'ISEA*, 1964, G20/150 Suppl., 227–35

———, *The Market-Planned Economy o Yugoslavia*, Minneapolis 1966.

J. Pelicon, "Sumarna ocena i neki problem privredne suradnje SFRJ sa zemljama u razvoju u 1966-1967. godini," in *Privredn odnosi Jugoslavije sa zemljama u razvoju* Ljubljana 1968, pp. 1–23.

M. Perović, "Još o prelaznom periodu," *Ekonomski pregled*, 1953, vol. IV, 29–42.

A. Perić, *Finansijska teorija i politika*, Beograd 1964.

I. Perišin, "Stabilizacija i monetarno-kreditna politika," *Ekonomist*, 1967, vol. XX, 103-120.

———, *Monetarno-kreditna politika*, Zagreb 1968.

———, "Antiinflatorna politika Jugoslavije poslije reforme," *Ekonomski pregled*, 1969, vol. XX, 497–530.

V. Pertot, *Yugoslav Foreign Trade*, Beograd 1960.

———, "Stabilizacija u uslovima disparitetnih odnosa troškova proizvodnje," *Ekonomist*, 1966, vol. XIX, 316–44.

M. Petrović, *Formiranje prihoda društveno-političkih zajednica u SR Srbiji i njihova raspodela izmedju republika, pokrajina i opština*, Beograd 1968.

J. Pokorn, "Razvoj našeg finansijskog sistema," *Finansije*, 1956, vol. XI, 1-10.

S. Popov, "Kretanje produktivnosti rada i

ličnih dohodaka u pojedinim granama u periodu od 1952 do 1966. godine," in *Obračuni i raspodela osobnih dohodaka u radnim organizacijama*, Zagreb 1968, pp. 613–33.

Z. Popov, "Osvrt na kretanje privrednog razvoja u svetu," *Ekonomska analiza*, 1968, vol. II, 353–65.

M. Popović, "O ekonomskim odnosima izmedju socijalističkih država," *Komunist*, 1949, *4*, 89–146.

———, "O sistemu ekonomske i socijalističke demokratije u Jugoslaviji," *Komunist*, 1952, *3–4*, 1–14.

———, *Društveno ekonomski sistem*, Beograd 1964.

S. Popović, "Merenje dohotka i njegova raspodela," *Ekonomska misao*, 1968, vol. I, 424–36.

E. Pusić, *Samoupravljanje*, Zagreb 1968.

R. Radovanović, *Poreski sistem FNRJ*, Beograd 1953.

———, "Budžet u toku proteklih deset godina," in *Razvoj privrede FNRJ*, Beograd, 1956a, pp. 443–51.

———, *Oporezivanje dohotka privrednih poduzeča*, Beograd 1956b.

———, "Budgetary System and Budget Expenditure," *Yugoslav Survey*, 1962, vol. III, 1111–22.

M. Radulović, *Sistem i politika cijena u Jugoslaviji*. Unpublished doctoral dissertation, Titograd 1968.

V. Rajković, "Ocjena ostvarivanja privredne reforme i aktuelni problemi," in *Aktuelni problemi ekonomske politike Jugoslavije*, Zagreb, 1969/1970, pp. 21–48.

V. Rašković, "Osnovni idejni i politički problemi ličnog rada u sistemu društvenog samoupravljanja," in *Privatni rad: Za ili protiv*, Beograd, 1967a.

———, *Društveno samoupravljanje i raspodela prema radu u Jugoslaviji*, Beograd 1967b.

M. Samardžija, "Metodološke i društvene osnove teorije raspodele dohodka," *Gledišta*, 1968, vol. XIX, 124–49, 293–304.

J. A. Schumpeter, *Capitalism, Socialism, and Democracy*, New York 1950.

B. Šefer, "Tržište u posleratnom periodu," in *Razvoj privrede FNRJ*, Beograd 1956.

———, "Problemi i politika razvoja lične i društvene potrošnje," in J. Sirotković, ed., *Suvremeni problemi jugoslavenske privrede i ekonomska politika*, Zagreb 1965.

———, "Rasponi ličnih dohodaka, njihovo formiranje i tendencije," in *Obračun i raspodela osobnih dohodaka u radnim organizacijama*, Zagreb 1968a, pp. 421–38.

———, *Ekonomski razvoj Jugoslavije i privredna reforma*, Beograd 1968b.

M. Sekulić, *Primjena strukturnih modela u planiranju privrednog razvoja*, Zagreb 1968.

P. Sicherl, *et al.*, *Izučavanje problema dopunskih sredstava republikama na trajnijoj osnovi*. Mimeograph, Institute of Economic Studies, Beograd 1968.

———, "Analiza nekih elemanata za ocenu stepena razvijenosti republika i pokrajina," *Ekonomska analiza*, 1969, vol. III, 5–28.

J. Sirotković, *Planiranje proširene reprodukcije u socijalizmu*, Zagreb 1951.

———, *Problemi privrednog planiranja u Jugoslaviji*, Zagreb 1961.

———, *Planiranje u sistemu samoupravljanja*, Zagreb 1966.

B. Šoškić, "Rast proizvodnje i zaposlenosti i mere ekonomske politike," *Ekonomist*, 1969a, vol. XXII, 143–55.

———, "Povećanje zaposlenosti u našem sistemu tržišne privrede," *Ekonomska misao*, 1969b, vol. II, 79–92.

S. Srdar, *Da li FNR Jugoslavija postaje agrarnouvozna i industrijski izvozna zemlja*, Zagreb 1953.

V. Stanovčić and A. Stojanović, eds., Books I and II, *Birokratija i tehnokratija*, Beograd 1966.

R. M. Stevanović, *Novčani i kreditni sistem*, Beograd 1954.

S. Stojanović, "Etatistički mit socijalizma," *praxis*, 1967, vol. III, 30–38.

R. Stojanović, "Stopa rasta socijalističke privrede," in R. Stojanović, ed., *Savremeni problemi privrednog razvoja u socijalizmu*, Beograd 1960.

M. Sukijasovič and Dj. Vujačić, *Industrial Cooperation and Joint Investment Ventures Between Yugoslav and Foreign Firms*, Beograd 1968.

P. Sweezy, "The Transition from Socialism to Capitalism?", *Monthly Review*, 1964, vol. 16, 569–90.

Z. Tanić, ed., *Radničko samoupravljanje; razvoj i problemi*, Beograd 1963.
T. Tišma, *Javne finansije*, Zagreb 1964.
M. Todorović, *Oslobodjenje rada*, Beograd 1965.
T. Tomić, "Dosadašnji razvoj raspodele ličnih dohodaka u SFRJ," in *Obračun i raspodela osobnih dohodaka u radnim organizacijama*, Zagreb 1968, pp. 3–22.
M. Toroman, "Oblici društvene svojine." Paper presented at the Symposium on Social Ownership, Serbian Academy of Science and Art, Beograd, Sept. 20–22, 1965.
M. Trklja, *Kamata na fondove u privredi*. Mimeograph, Institute of Economic, Zagreb 1968.
S. Turčinović, "Financing Socio-Political Units," *Yugoslav Survey*, 1968, *2*, 59–74.
R. Uvalić, "Zakon vrednosti i njegovo korišćenje u planiranju narodne privrede," *Ekonomist*, 1948, *1*, 20–27.
———, "O nekim principima našeg privrednog sistema i problemi njihove primene," *Ekonomist*, 1954, *3*, 5–17.
———, "Funkcije tržišta i plana u socijalističkoj privredi," *Ekonomist*, 1962, vol. XV, 205–19.
———, "Trojna ekonomska suradnja Jugoslavije, Indije i UAR–a," in *Privredni odnosi Jugoslavije sa zemljama u razvoju*, Ljubljana 1968, pp. 128–46.
F. Vasić, "Investment in the Post–War Period," *Yugoslav Survey*, 1963, *15*, 2153–172.
Z. Vidaković, *Promene u strukturi jugoslavenskog društva i Savez Komunista*, Beograd 1967.
D. Vojnić, "Investiciona politika i sistem proširene reprodukcije," in *Aktuelni problemi ekonomske politike Jugoslavije 1969/1970*, Zagreb 1969, pp. 75–92.
M. Vučković, *Naš novi planski i finansijski sistem*, Beograd 1952.
———, "Preduzeće i kredit," *Ekonomski anali*, 1956, vol. II, 166–85.
———, *Kreditni sistem u FNRJ*, Beograd 1957.
———, "The Recent Development of the Money and Banking System of Yugoslavia," *Journal of Political Economy*, 1963, vol. LXXI, 363–77.
———, "Dosadašnja inflaciona kretanja u Jugoslaviji," *Ekonomist*, 1967, vol. XX 121–140.
D. Vuković, "Price Formation and Social Price Control," *Yugoslav Survey*, 1968, *1* 51–58.
R. Vuksanović, "Credit and Money," *Yugoslav Survey*, 1966, vol. VII, 3461–74.
H. M. Wachtel, *Workers' Management and Wage Differentials in Yugoslavia*. Unpublished doctoral dissertation, Univ. Mich. 1969.
B. Ward, "Workers' Management in Yugoslavia," *Journal of Political Economy*, 1957, vol. LXV, 373–86.
———, "The Firm in Illyria: Market Syndicalism," *American Economic Review*, 1958, vol. XLVIII, 566–89.
———, "Marxism-Horvatism: A Yugoslav Theory of Socialism," *American Economic Review*, 1967, vol. LVII, 509–23.
M. Žiberna, "Neki problemi ekonomskih odnosa s evropskom ekonomskom zajednicom," *Medunarodni problemi*, 1969, vol. XXI, 51–66.
J. Županov, "Radni kolektiv i ekonomska jedinica u svjetlu organizacione teorije," *Ekonomski pregled*, 1962, vol. XIII, 143–69.
———, *O problemima upravljanja i rukovodjenja u radnoj organizaciji*, Zagreb 1967a.
———, "Proizvodjač i riziko—Neki socijalno–psihološki aspekti kolektivnog poduzetništva," *Ekonomist*, 1967b, vol. XX, 389–408.
Ekonomist, "Diskusija ekonomista o prednacrtu ustava," 1962, vol. XV, 439–517.
Finansijski institut, *Finansijski sistem FNR Jugoslavije*, Beograd 1949.
Informativni jiručnik o Jugoslaviji, "Izvještaji savezne planske komisije," October 1948.
Institut društvenih nauka, *Koncepcija i verifikacija specifične cene priozvodnje u jugoslavenskoj privredi 1964. i 1965*, Beograd 1968.
Institut ekonomskih nauka, *Nauka i ekonomska politika*, Beograd 1968a.
———, *Sumarna analiza privrednih kretanja i prijedlozi za ekonomsku politiku*, Beograd 1968b.
———, *Ocjena ekonomske sitiacije i predvidjanja daljeg razvoja*, Beograd 1969.

nstitut za spoljnu trgovinu, *Analiza devizne reforme iz 1961*, Beograd 1964.

nternational Labour Office, *Workers' Management in Yugoslavia*, Geneva 1962.

ugoslavenski institut za ekonomska istraživanja, *Sumarna analiza privrednih kretanja i prijedlozi za ekonomsku politiku*, Beograd 1968.

Narodna banka Jugoslavije. *Novčano-kreditna politika i stabilnost dinara*. Mimeographed paper, Beograd 1965.

Program Saveza Komunista Jugoslavije, Beograd 1958.

Savezna skupština, *Osnovni problemi daljneg razvoja privrednog sistema*, Beograd 1964.

———, *Privredne reforme*, Beograd 1965.

———, *Osnove sistema društvenog planiranja*, Beograd 1966a.

———, *Devizni i spoljnotrgovinski režim*, Beograd 1966b.

Savezni zavod za statistiku, *Jugoslavija 1945–1964*, Beograd 1965.

Savjetovanje jugoslavenskih ekonomista, Zagreb 17–19 januara 1963, "Aktuelni problemi privrednog razvoja i privrednog sistema Jugoslavije," *Ekonomist*, 1963, *1*.

Savjetovanje jugoslovenskih ekonomista, "Problemi teorije i politika cena," *Ekonomist*, 1964, vol. XVII, 499–792.

———, Ljubljana, 9–11 marta 1967, "O uslovima stabilizacije jugoslavenske privrede," *Ekonomist*, 1966, *1–4*, 1967, *1–2*.

Simpozij jugoslavenskih–čekoslovačkih filosofa, "Savremeni trenutak socijalizma," *Filosofija*, 1969, vol. XIII, *2*, 1–98.

Yugoslav Survey, "Resolution of Federal Assembly on the Guidelines for Drawing up Yugoslavia's Social Plan for the 1964–1970 Period," 1964, 2703–16.

U. S. Congress, Senate Subcommittee on Antitrust and Monopoly of the Committee on the Judiciary, *Problems of Market Power and Public Policy in Yugoslavia*, by J. Dirlam, 90th Cong., 2nd sess., 1968, pp. 3758–85.